PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C12Q 1/68, G01N 33/53, C12P 19/34, C12N 5/10, 1/21, C07K 5/00, 14/00, 16/00, C07H 21/02, 21/04

A1

(11) International Publication Number: WO 96/25519

(43) International Publication Date: 22 August 1996 (22.08.96)

(21) International Application Number: PCT/US96/01938

(22) International Filing Date: 15 February 1996 (15.02.96)

(30) Priority Data:
08/390,878 17 February 1995 (17.02.95) US

(71) Applicant: PATHOGENESIS CORPORATION [US/US]; Suite 150, 201 Elliott Avenue West, Seattle, WA 98119 (US).

(72) Inventors: STOVER, Charles, Kendall; 7640 81st Place S.E., Mercer Island, WA 98040 (US). MAHAIRAS, Gregory, G.; 3312 39th West, Seattle, WA 98199 (US).

(74) Agents: HUNTER, Tom et al.; Townsend and Townsend and Crew, Steuart Street Tower, One Market, San Francisco, CA 94105-1492 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BB, BG, BR, BY, CA, CH, CN, CZ, DE, DK, EE, ES, FI, GB, GE, HU, IS, JP, KE, KG, KP, KR, KZ, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, TJ, TM, TR, TT, UA, UG, UZ, VN, ARIPO patent (KE, LS, MW, SD, SZ, UG), Eurasian patent (AZ, BY, KG, KZ, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: VIRULENCE-ATTENUATING GENETIC DELETIONS

(57) Abstract

The present invention provides specific genetic deletions that result in an avirulent phenotype of a mycobacterium. These deletions may be used as phenotypic markers of providing a means for distinguishing between disease-producing and non-disease producing mycobacteria.

IDS

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia	GB	United Kingdom	MW	Malawi
AT	Austria	GE	Georgia	MX	Mexico
AU	Australia	GN	Guinea	NE	Niger
BB	Barbados	GR	Greece	NL	Netherlands
BE	Belgium	HU	Hungary	NO	Norway
BF	Burkina Faso	IE	Ireland	NZ	New Zealand
BG	Bulgaria	IT	Italy	PL	Poland
BJ	Benin	JP	Japan	PT	Portugal
BR	Brazil	KE	Kenya	RO	Romania
BY	Belarus	KG	Kyrgystan	RU	Russian Federation
CA	Canada	KP	Democratic People's Republic of Korea	SD	Sudan
CF	Central African Republic			SE	Sweden
CG	Congo	KR	Republic of Korea	SG	Singapore
CH	Switzerland	KZ	Kazakhstan	SI	Slovenia
CI	Côte d'Ivoire	LI	Liechtenstein	SK	Slovakia
CM	Cameroon	LK	Sri Lanka	SN	Senegal
CN	China	LR	Liberia	SZ	Swaziland
CS	Czechoslovakia	LT	Lithuania	TD	Chad
CZ	Czech Republic	LU	Luxembourg	TG	Togo
DE	Germany	LV	Latvia	TJ	Tajikistan
DK	Denmark	MC	Monaco	TT	Trinidad and Tobago
EE	Estonia	MD	Republic of Moldova	UA	Ukraine
ES	Spain	MG	Madagascar	UG	Uganda
FI	Finland	ML	Mali	US	United States of America
FR	France	MN	Mongolia	UZ	Uzbekistan
GA	Gabon	MR	Mauritania	VN	Viet Nam

VIRULENCE-ATTENUATING GENETIC DELETIONS

BACKGROUND OF THE INVENTION

Mycobacterium tuberculosis (MTB) infects over ten million people each year and kills over three million, making it the infectious agent causing the greatest mortality worldwide. In an effort to combat *Mycobacterium tuberculosis*, vaccination programs using a viable attenuated strain of *Mycobacterium bovis* called bacille Calmette-Guérin (BCG) have been established in more than 120 countries over the course of the last 5 decades. Although widely used and considered safe enough to administer to infants, the BCG vaccine is controversial for two principle reasons: 1) Efficacy for BCG vaccines against tuberculosis has varied from 0-85% in different clinical trials; and 2) Immunization with BCG sensitizes vaccinees to the tubercular antigens used in the tuberculin skin test, confounding attempts to discriminate between BCG immunization and TB infection. For these two reasons, especially the latter, BCG is not used in the United States where surveillance with the tuberculin test is preferred.

The original Pasteur BCG strain was developed by multiple (230 times) serial passages in liquid culture. BCG has never been shown to revert to virulence in animals indicating that the attenuating mutations in BCG are stable deletions and/or multiple mutations which cannot revert. However, the mutations which arose during serial passage of the original BCG strain have never been identified. Moreover, recent efforts to genetically complement BCG virulence with genomic libraries of virulent tubercle bacilli have also been unsuccessful again suggesting that multiple unlinked mutations are responsible for the attenuation of BCG virulence. The antigenicity of BCG and the characteristics leading to its avirulence are thus poorly understood.

SUMMARY OF THE INVENTION

The present invention provides specific genetic deletions that account for the avirulent phenotype of the bacille Calmette-Guérin (BCG) strain of *Mycobacterium bovis*. These deletions may be used as phenotypic markers of providing a means for distinguishing between disease-producing and non-disease producing mycobacteria.

In a preferred embodiment, this invention provides for nucleic acid sequences that are markers for avirulent or virulent mycobacteria. The sequences uniquely characterize the presence or absence of deletions that result in an avirulent phenotype. More specifically the sequence are either deletion junction sequence or deletion sequences or subsequences within deletion junction sequences or deletion sequences. Thus, this invention provides for a marker for an avirulent mycobacterium comprising a first nucleic acid that hybridizes under stringent conditions with a second nucleic acid or a complement of the second nucleic acid where the second nucleic acid or its complement includes BCGΔ1a, BCGΔ1b, BCGΔ2a, BCGΔ2b, BCGΔ3a, BCGΔ3b, BCGΔ1ab, BCGΔ2ab, BCGΔ3ab, BCGΔ1, BCGΔ2, and BCGΔ3. In a particularly preferred embodiment, the marker specifically hybridizes under stringent conditions to a nucleic acid from BCG, but not to a nucleic acid from *Mycobacterium tuberculosis* or *Mycobacterium bovis*, or alternatively, the marker specifically hybridizes under stringent conditions to a nucleic acid from *Mycobacterium tuberculosis* or *Mycobacterium bovis*, but not to a nucleic acid from BCG. The marker may be the full length BCGΔ1a, BCGΔ1b, BCGΔ2a, BCGΔ2b, BCGΔ3a, BCGΔ3b, BCGΔ1ab, BCGΔ2ab, BCGΔ3ab, BCGΔ1, BCGΔ2, and BCGΔ3 or a subsequence within any of these regions. The marker may also include a nucleic acid having at least 80%, preferably 90%, more preferably 95%, and most preferably 98% percent sequence identity with BCGΔ1a, BCGΔ1b, BCGΔ2a, BCGΔ2b, BCGΔ3a, BCGΔ3b, BCGΔ1ab, BCGΔ2ab, BCGΔ3ab, BCGΔ1, BCGΔ2, or BCGΔ3. The marker may also include a sequence selected from an open reading frame of a the deletion sequences BCGΔ1, BCGΔ2, BCGΔ3. Suitable open reading frames are indicated in Figures 4, 5, and 6.

The above described marker may be a probe. The probe may be labeled by a number of means including, but not limited to radioactive, fluorescent, enzymatic, and colorimetric labels.

In another embodiment, this invention provides for polypeptides encoded by a subsequence of the BCGΔ1, BCGΔ2, or BCGΔ3 deletions. In particular, the subsequence may be selected from an open reading frame (ORF) present in one of these deletion sequences. This invention also provides for monoclonal or polyclonal antibodies that

specifically bind polypeptides encoded by one or more subsequences of the BCGΔ1, BCGΔ2, or BCGΔ3 deletions.

In still another embodiment, this invention provides for a recombinant cell comprising a first nucleic acid that hybridizes under stringent conditions with a second nucleic acid or a complement of the second nucleic acid where the second nucleic acid or its complement is BCGΔ1a, BCGΔ1b, BCGΔ2a, BCGΔ2b, BCGΔ3a, BCGΔ3b, BCGΔ1ab, BCGΔ2ab, BCGΔ3ab, BCGΔ1, BCGΔ2, or BCGΔ3. The recombinant cell may be a mycobacterium. The recombinant cell may express a polypeptide encoded by any of BCGΔ1a, BCGΔ1b, BCGΔ2a, BCGΔ2b, BCGΔ3a, BCGΔ3b, BCGΔ1ab, BCGΔ2ab, BCGΔ3ab, BCGΔ1, BCGΔ2, and BCGΔ3. More preferably, the recombinant cell expresses a polypeptide encoded by an intact open reading frame present in any of these regions. The cell may also be a mycobacterium having one or more deletions in the BCGΔ1, BCGΔ2, or BCGΔ3 genomic regions where the deletions result in the attenuation of an otherwise virulent strain of mycobacterium and wherein the deletions are present in up to two of the genomic regions.

In still yet another embodiment, this invention provides a method of distinguishing between an attenuated and a virulent mycobacterium. The method involves detecting the presence or absence of a first nucleic acid that hybridizes under stringent conditions with a second nucleic acid or a complement of the second nucleic acid where the second nucleic acid or its complement is BCGΔ1a, BCGΔ1b, BCGΔ2a, BCGΔ2b, BCGΔ3a, BCGΔ3b, BCGΔ1ab, BCGΔ2ab, BCGΔ3ab, BCGΔ1, BCGΔ2, or BCGΔ3. The first nucleic acid may include any of the markers described above. A particularly preferred marker is one that specifically hybridizes under stringent conditions to a nucleic acid from BCG, but not to a nucleic acid from *Mycobacterium tuberculosis* or *Mycobacterium bovis*, or alternatively, that specifically hybridizes under stringent conditions to a nucleic acid from *Mycobacterium tuberculosis* or *Mycobacterium bovis*, but not to a nucleic acid from BCG. The method may involve amplifying either the first nucleic acid by any of a number of methods including, for example, polymerase chain reaction. The detection may involve detecting the first nucleic acid, for example, as in a Southern blot, or alternatively, detecting a polypeptide encoded by the first nucleic acid. More specifically, the polypeptide may be

a encoded by an open reading frame (ORF) selected from BCGΔ1, BCGΔ2, or BCGΔ3. The polypeptide may be visualized by a number of means well known to those of skill in the art including antibody hybridization such as direct or indirect binding of labeled antibody.

This invention additionally provides a method for determining whether an attenuated or a virulent Mycobacterium is present in a sample. This method involves providing a first nucleic acid that hybridizes under stringent conditions with a second nucleic acid or a complement of the second nucleic acid where the second nucleic acid or its complement is BCGΔ1a, BCGΔ1b, BCGΔ2a, BCGΔ2b, BCGΔ3a, BCGΔ3b, BCGΔ1ab, BCGΔ2ab, BCGΔ3ab, BCGΔ1, BCGΔ2, or BCGΔ3; and hybridizing the first nucleic acid to the biological sample. The first nucleic acid may include any of the markers described above. A particularly preferred marker is one that specifically hybridizes under stringent conditions to a nucleic acid from BCG, but not to a nucleic acid from *Mycobacterium tuberculosis* or *Mycobacterium bovis*, or alternatively, that specifically hybridizes under stringent conditions to a nucleic acid from *Mycobacterium tuberculosis* or *Mycobacterium bovis*, but not to a nucleic acid from BCG. The method may involve amplifying either the first nucleic acid by any of a number of methods including, for example, polymerase chain reaction. The detection may involve detecting the first nucleic acid, for example, as in a Southern blot, or alternatively, detecting a polypeptide encoded by the first nucleic acid. More specifically, the polypeptide may be a encoded by an open reading frame (ORF) selected from BCGΔ1, BCGΔ2, or BCGΔ3. The method may also include detecting the hybridized first nucleic acid. This may involve direct detection of a label or additionally involve an amplification step and subsequent detection of the amplified product.

Finally, this invention provides a method of producing an attenuated-virulence mycobacterium. This method involves deleting from the genomic DNA of a virulent mycobacterium a first nucleic acid that specifically hybridizes under stringent conditions with a second nucleic acid or a complement of said second nucleic acid where said second nucleic acid or complement of said second nucleic acid is selected from the group consisting of BCGΔ1, BCGΔ2, and BCGΔ3. The first nucleic acid may be BCGΔ1, BCGΔ2, or BCGΔ3, or alternatively, it may be a promoter, other control element or an open reading frame from BCGΔ1, BCGΔ2, or BCGΔ3.

Definitions

Although any methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, the preferred methods and materials are described. For purposes of the present invention, the following terms are defined below.

The phrase "specifically detect" as used herein refers to the process of determining that a particular subsequence is present in a DNA sample. A DNA sequence may be specifically detected through a number of means known to those of skill in the art. These would include, but are not limited to amplification of the particular target sequence through polymerase chain reaction or ligase chain reaction, hybridization of the sequence to a labeled probe, and binding by labelled ligands or monoclonal antibodies. For a discussion of various means of detection of specific nucleic acid sequences see Perbal, B. *A Practical Guide to Molecular Cloning*, 2nd Ed. John Wiley & Sons, N.Y. (1988) which is incorporated herein by reference.

The phrase "select subsequence" is used herein to refer to a particular DNA subsequence that is of interest. It is often a predetermined or known sequence of nucleic acid bases. A select subsequence is typically chosen because of a unique sequence identity. Typically a select subsequence is targeted for DNA amplification and often is useful as a specific marker for the presence of a particular gene or a deletion of a particular nucleic acid sequence.

The term "oligonucleotide" refers to a molecule comprised of two or more deoxyribonucleotides or ribonucleotides. Oligonucleotides may include, but are not limited to, primers, probes, nucleic acid fragments to be detected, and nucleic acid controls. Oligonucleotides include naturally occurring nucleotides, chemically modified naturally occurring nucleotides and synthetic nucleotides. The exact size of an oligonucleotide depends on many factors and the ultimate function or use of the oligonucleotide.

The term "primer" refers to an oligonucleotide, whether natural or synthetic, capable of acting as a point of initiation of DNA synthesis under conditions in which synthesis of a primer extension product complementary to a nucleic acid strand is induced, *i.e.*, in the presence of four different nucleoside triphosphates and an agent for polymerization (*i.e.*, DNA polymerase or reverse transcriptase) in an appropriate buffer and at a suitable temperature. A primer is preferably a single-stranded oligodeoxyribonucleotide.

The appropriate length of a primer depends on the intended use of the primer but typically ranges from 15 to 25 nucleotides. Short primer molecules generally require cooler temperatures to form sufficiently stable hybrid complexes with the template. A primer need not reflect the exact sequence of the template but must be sufficiently complementary to hybridize with a template.

The phrase "PCR primers competent to amplify" as used herein refers to a pair of PCR primers whose sequences are complementary to DNA subsequences immediately flanking the DNA subsequence (target sequence) which it is desired to amplify. The primers are chosen to bind specifically those particular flanking subsequences and no other sequences present in the sample. The PCR primers are thus preferably chosen to amplify the unique target sequence and no other. Alternatively, the PCR primers may be selected to bind to sequences other than the target sequence where the amplification products can be subsequently distinguished (*e.g.* where the desired amplified sequence is different in size than other amplified sequences).

"Amplifying" or "amplification", which typically refer to an "exponential" increase in target nucleic acid, are used herein to describe both linear and exponential increases in the number of a select target sequence of nucleic acid.

The term "antisense orientation" refers to the orientation of nucleic acid sequence from a structural gene that is inserted in an expression cassette in an inverted manner with respect to its naturally occurring orientation. When the sequence is double stranded, the strand that is the template strand in the naturally occurring orientation becomes the coding strand, and vice versa.

The term "deletion" refers to a region of a nucleic acid which is not present in an organism, but which is present in another related organism. In the context of mycobacteria, a deletion refers, *e.g.*, to a region of nucleic acid which is not present in one strain of mycobacteria, but which is present in another related strain. For instance, an avirulent mycobacterial strain can have a deletion in its genome relative to the genome of a related virulent mycobacterial strain.

The term "deletion junction" refers to the region of a nucleic acid spanning the insertion point of a deletion. Thus, where a region of a nucleic acid sequence is deleted (*i.e.* a deletion is present), the deletion junction spans the nucleotides that are immediately adjacent to the deletion. Conversely, where a region of a nucleic acid sequence is not

deleted (*i.e.* the deletion is absent), two deletion junctions are present, each spanning respectively one end of the deletion sequence and its flanking sequence.

The following terms are used to describe the sequence relationships between two or more polynucleotides: "reference sequence", "comparison window", "sequence identity", "percentage of sequence identity", and "substantial identity". A "reference sequence" is a defined sequence used as a basis for a sequence comparison; a reference sequence may be a subset of a larger sequence, for example, as a segment of a full-length cDNA or gene sequence given in a sequence listing, such as a polynucleotide sequence of Figures 1, 2, or 3, or may comprise a complete cDNA or gene sequence.

Generally, a reference sequence is at least 10 nucleotides in length, frequently at least 20 to 25 nucleotides in length, and often at least 50 nucleotides in length. Sequence comparisons between two (or more) polynucleotides are typically performed by comparing sequences of the two polynucleotides over a "comparison window" to identify and compare local regions of sequence similarity. A "comparison window", as used herein, refers to a segment of at least 10 contiguous nucleotide positions wherein a polynucleotide sequence may be compared to a reference sequence of at least 10 contiguous nucleotides and wherein the portion of the polynucleotide sequence in the comparison window may comprise additions or deletions (*i.e.*, gaps) of 20 percent or less as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences.

Optimal alignment of sequences for aligning a comparison window may be conducted by the local homology algorithm of Smith and Waterman *Adv. Appl. Math.* 2: 482 (1981), by the homology alignment algorithm of Needleman and Wunsch *J. Mol. Biol.* 48: 443 (1970); by the search for similarity method of Pearson and Lipman *Proc. Natl. Acad. Sci. (USA)* 85: 2444 (1988), or by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package Release 7.0, Genetics Computer Group, 575 Science Dr., Madison, WI), or by inspection, and the best alignment (*i.e.*, resulting in the highest percentage of sequence similarity over the comparison window) generated by the various methods is selected.

The term "sequence identity" means that two polynucleotide sequences are identical (*i.e.*, on a nucleotide-by-nucleotide basis) over the window of comparison. The term "percentage of sequence identity" is calculated by comparing two optimally aligned

sequences over the window of comparison, determining the number of positions at which the identical nucleic acid base (*e.g.*, A, T, C, G, U, or I) occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison (*i.e.*, the window size), and multiplying the result by 100 to yield the percentage of sequence identity. The term "identical" in the context of two nucleic acid or polypeptide sequences refers to the residues in the two sequences which are the same when aligned for maximum correspondence.

The terms "isolated" or "biologically pure" refer to material which is substantially or essentially free from components which normally accompany it as found in its native state. The isolated nucleic acid probes of this invention do not contain materials normally associated with their *in situ* environment, in particular nuclear, cytosolic or membrane associated proteins or nucleic acids other than those nucleic acids intended to comprise the nucleic acid probe itself.

The term "marker" refers to a characteristic which distinguishes one class of cells or compositions from a second class of cells or compositions. For instance, the deletions and deletion junctions described herein can be used to distinguish between strains (*e.g.*, virulent and avirulent strains) of mycobacteria. While markers are indicators of associated features or properties, as used herein, markers may also be used for purposes other than indicating the associated feature or property. Thus, for example, a nucleic acid marker of virulence identifies a particular nucleic acid which may be used in a variety of contexts other than simply indicating virulence.

The term "nucleic acid" refers to a deoxyribonucleotide or ribonucleotide polymer in either single- or double-stranded form, and unless otherwise limited, encompassing known analogues of natural nucleotides that can function in a similar manner as naturally occurring nucleotides. Unless otherwise indicated, a particular nucleic acid sequence includes the complementary sequence thereof.

The term "operably linked" refers to functional linkage between a promoter and a second sequence, wherein the promoter sequence initiates transcription of RNA corresponding to the second sequence.

The term "peptide" or "polypeptide" refers to an amino acid polymer which is encoded by a nucleic acid. The peptide or polypeptide may include naturally occurring or modified amino acids.

The terms "probe" or "nucleic acid probe" refer to a molecule that binds to a specific sequence or subsequence of a nucleic acid. A probe is preferably a nucleic acid which binds through complementary base pairing to the full sequence or to a subsequence of a target nucleic acid. It will be understood by one of skill in the art that probes may bind target sequences lacking complete complementarily with the probe sequence depending upon the stringency of the hybridization conditions. The probes are preferably directly labelled as with isotopes, chromophores, lumiphores, chromogens, or indirectly labelled such with, *e.g.*, biotin to which a streptavidin complex may later bind. By assaying for the presence or absence of the probe, one can detect the presence or absence of the selected sequence or subsequence.

The term "labeled nucleic acid probe" refers to a nucleic acid probe that is bound, either covalently, through a linker, or through ionic, van der Waals or hydrogen "bonds" to a label such that the presence of the probe may be detected by detecting the presence of the label bound to the probe.

The term "recombinant" when used with reference to a cell indicates that the cell replicates or expresses a nucleic acid, or expresses a peptide or protein encoded by DNA whose origin is exogenous to the cell. Recombinant cells can express genes that are not found within the native (non-recombinant) form of the cell. Recombinant cells can also express genes found in the native form of the cell wherein the genes are re-introduced into the cell by artificial means.

The term "sample" refers to a material with which bacteria may be associated. Frequently the sample will be a "clinical sample" which is a sample derived from a patient. Such samples include, but are not limited to, sputum, blood, blood cells (*e.g.*, white cells), tissue or fine needle biopsy samples, urine, peritoneal fluid, and pleural fluid, or cells therefrom. It will be recognized that the term "sample" also includes supernatant from eukaryotic cell cultures (which may contain free bacteria), cells from cell or tissue culture, and other media in which it may be desirable to detect mycobacteria (*e.g.*, food and water).

The term "subsequence" in the context of a particular nucleic acid sequence refers to a region of the nucleic acid equal to or smaller than the specified nucleic acid.

The term "substantial identity" or "substantial similarity" indicates that a nucleic acid or polypeptide comprises a sequence that has at least 90% sequence identity to a reference sequence, or preferably 95%, or more preferably 98% sequence identity to the

reference sequence, over a comparison window of at least about 10 to about 100 nucleotides or amino acid residues. An indication that two polypeptide sequences are substantially identical is that one protein is immunologically reactive with antibodies raised against the second protein. An indication that two nucleic acid sequences are substantially identical is that the polypeptides which the first nucleic acids encodes is immunologically cross reactive with the polypeptide encoded by the second nucleic acid.

Another indication that two nucleic acid sequences are substantially identical is that the two molecules hybridize to each other under stringent conditions. Stringent conditions are sequence-dependent and will be different with different environmental parameters. Generally, stringent conditions are selected to be about 5°C to 20°C lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength and pH. The T_m is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched probe. Typically, stringent conditions will be those in which the salt concentration is at least about 0.2 molar at pH 7 and the temperature is at least about 60°C.

The term "uninterrupted reading frame" or "open reading frame" refers to a DNA sequence (*e.g.*, cDNA) lacking a stop codon or other intervening, untranslated sequence. An intact open reading frame refers to a full length uninterrupted reading frame or minor variations thereof.

The term "virulent" in the context of mycobacteria refers to a bacterium or strain of bacteria that replicates within a host cell or animal at a rate that is detrimental to the cell or animal within its host range. More particularly virulent mycobacteria persist longer in a host than avirulent mycobacteria. Virulent mycobacteria are typically disease producing and infection leads to various disease states including fulminant disease in the lung, disseminated systemic milliary tuberculosis, tuberculosis meningitis, and tuberculosis abscesses of various tissues. Infection by virulent mycobacteria often results in death of the host organism. Typically, infection of guinea pigs is used as an assay for mycobacterial virulence. In contrast, the term "avirulent" refers to a bacterium or strain of bacteria that either does not replicate within a host cell or animal within its host range, or replicates at a rate that is not significantly detrimental to the cell or animal.

The term BCG-like avirulence, as used herein refers to an attenuated virulence brought about by one of the deletions of the present invention.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 shows the complete sequence listing of the BCG deletion region 1 including flanking sequences. The deletion, designated BCGΔ1, is located between nucleotide 2327 and nucleotide 11126.

Figure 2 shows the complete sequence listing of the BCG deletion region 2 including flanking sequences. The deletion, designated BCGΔ2, is located between nucleotide 3382 and nucleotide 14071.

Figure 3 shows the complete sequence listing of the BCG deletion region 3 including flanking sequences. The deletion, designated BCGΔ3, is located between nucleotide 1406 and nucleotide 10673. "N" represents "A", "C", "G", or "T".

Figure 4 shows a map of the deletion sequence BCGΔ1. This map identifies the various open reading frames (ORFs) and indicates their location within the deletion sequence. Ribozome binding sites and homologies to the predicted encoded proteins are shown.

Figure 5 shows a map of the deletion sequence BCGΔ2. This map identifies the various open reading frames (ORFs) and indicates their location within the deletion sequence. Ribozomal binding sites and homologies to the predicted encoded proteins are shown.

Figure 6 shows a map of the deletion sequence BCGΔ3. This map identifies the various open reading frames (ORFs) and indicates their location within the deletion sequence. Ribozome binding sites and homologies to the predicted encoded proteins are shown. The sequence of a small region, estimated to be much less than 200 bp and located close to 9400 bp in Figure 3, remains to be determined. Therefore, the base pair coordinates given in the region 3 map 3' to the 9kb marker are approximations. The precise sequence determination of this region is likely to effect the length of open reading frames 3H and 3L.

Figure 7 illustrates the deletion junction regions of BCGΔ1, BCGΔ2, and BCGΔ3. The "terminal" deletion junction regions formed by the flanking sequences and the terminal regions of the deletion sequences are identified as BCGΔ1a, BCGΔ1b, BCGΔ2a, BCGΔ2b, and BCGΔ3a, and BCGΔ3b. When the deletion is present (the deletion sequences

are missing) the respective "a" and "b" sequences will be juxtaposed, thereby forming deletion "spanning" junction sequences designated BCGΔ1ab, BCGΔ2ab, and BCGΔ3ab, respectively.

Figure 8 shows EcoRI and BamHI restricted chromosomal DNAs from *Mycobacterium bovis*, BCG Connaught, and *Mycobacterium tuberculosis* strains H37Ra, H37Rv, and Erdman probed with ^{32}P labeled BCG subtracted probe.

DETAILED DESCRIPTION

This invention reflects the discovery of genetic deletions in mycobacteria that result in an avirulent genotype such as is exhibited by the bacille Calmette-Guérin (BCG) mycobacterium. The original Pasteur bacille Calmette-Guérin (BCG) strain was developed by multiple (230 times) serial passages in liquid culture. BCG has never been shown to revert to virulence in animals indicating that the attenuating mutations in BCG are stable deletions and/or multiple mutations that cannot revert. The mutations that arose during serial passage of the original BCG strain were not previously known. Recent efforts to genetically complement BCG virulence with genomic libraries of virulent tubercle bacilli were unsuccessful, again suggesting that multiple unlinked mutations are responsible for the attenuation of BCG virulence.

The genetic deletions leading to the avirulent phenotype of BCG were identified by genomic subtractions between Connaught strain of BCG and MBV/MTB. The subtracted probe resulting from the genomic subtraction between BCG and the H37 Rv strain of *M. tuberculosis* was subsequently used to identify and clone three regions from a cosmid library of *Mycobacterium bovis* genomic DNA. Southern blot mapping and DNA sequence comparisons between BCG and *M. bovis* showed that three regions, designated regions 1-3, contained DNA segments of approximately 9 kb, 11 kb and 9 kb respectively, which are deleted in the Connaught strain of BCG. Precise deletion junctions were identified for each region by comparisons of BCG and corresponding virulent MBV sequences. The respective deletions, designated BCGΔ1, BCGΔ2 and BCGΔ3 are illustrated in Figures 1-3.

One of skill in the art will appreciate that the deletions encompassed by BCGΔ1, BCGΔ2 and BCGΔ3 may be utilized in a variety of contexts. For example, the deletions may be utilized to distinguish between avirulent and virulent strains of

mycobacteria thereby providing early detection of patients at risk for tuberculosis. This is of particular importance where mycobacteria are identified in a sample from a patient that has been previously vaccinated with BCG. In this context it may be critical to determine whether mycobacteria identified in a biological sample from such a patient are pathogenic.

In another embodiment, the preparation of mycobacteria containing the deletions of the present invention may provide superior vaccines to BCG which has long been known to have marginal efficacy. Thus, for example, a *Mycobacterium tuberculosis* may contain a full BCGΔ1 deletion or a smaller deletion within BCGΔ1 (*e.g.* one or more open reading frames) rendering it avirulent. An avirulent MTB will provide a more efficient vaccine because it is antigenically more similar to MTB than is BCG. Moreover, an MTB rendered avirulent by the production of smaller deletions within the deletion regions identified in this invention will present more antigenic determinants.

Since the loss of virulence is due to the loss of gene products expressed by the nucleic acid sequences comprising the deletion regions, the BCGΔ1, BCGΔ2 and BCGΔ3 deletion sequences and proteins encoded within these deletion sequences provide suitable targets for drug screening. Thus, the use of deleted sequences as targets to screen for drugs that inhibit or interfere with transcription, translation, or post-translational processing of proteins encoded by the deletion sequences, or with the deletion encoded polypeptides themselves, provides an assay for anti-mycobacterial agents. In particular, the use of reporter genes such as firefly luciferase (FFlux), ß-galactosidase (BGal), and the like, under the control of promoters present in the deletion sequence provide a rapid assay for drugs regulating activity originating in this region. Conversely, since the protein products of the deletion sequences are presumably expressed in virulent mycobacterial species, proteins expressed by deletion sequences may make good antigens for antimycobacterial vaccines.

Finally, as the viability of BCG demonstrates, deletion regions BCGΔ1, BCGΔ2 and BCGΔ3 are not required for mycobacterial growth and reproduction. Thus, these deletion regions provide good insertion points for the expression of heterologous DNA. The heterologous DNA sequences may be under the control of endogenous inducible or constitutive promoters typically found in the deletion sequences, or alternatively, they may be under the control of introduced promoters, either constitutive or inducible, exogenous to mycobacteria.

I. Detection of Deletions

As indicated above, the deletions identified in the present invention provide useful markers for the identification of an avirulent (or conversely a virulent) mycobacterial phenotype. Specifically, determination of avirulence simply requires the detection of the presence or absence of the deletion (either BCGΔ1, BCGΔ2, or BCGΔ3, or deletions within these regions). Where the deletion is present in the bacterial DNA, the bacterium expresses a BCG-like avirulent phenotype. Conversely, where the deletion is absent in the bacterial DNA, the bacterium does not express a BCG-like avirulence. While this may indicate that the bacterium is virulent, one of skill will appreciate that the bacterium may still be avirulent due to the presence of other mutations or deletions. Nevertheless, screening for the presence of the deletion provides a means of detecting a BCG-like avirulent mycobacterium.

Means of detecting deletions are well known to those of skill in the art. Generally, the deletions may be detected either by detecting the presence or absence of deletion junctions, or, alternatively, by detecting the presence or absence of the sequences contained within the deletion (deletion sequences). Where a nucleic acid sequence is deleted (*i.e.*, a deletion is present), the sequences that previously flanked the deleted sequence are juxtaposed, thereby forming a new deletion junction that spans the deletion. Detection of the presence of such a "spanning" deletion junction indicates the presence of the deletion and thus the avirulent phenotype.

Conversely, where the nucleic acid sequence is not deleted (the deletion is not present) the spanning junction sequence will be absent (See, *e.g.* Figure 7). The "terminal" deletion junction sequences flanking each endpoint of the deletion region are present and detection of these terminal deletion junctions indicates the absence of a deletion. Spanning deletion junction regions and terminal deletion junctions suitable for detecting the deletions of the present invention are illustrated in Figure 7 and in Table 1.

Table 1. Nucleic acid sequences comprising deletion junctions. The symbol "|" indicates the insertion point of the deletion sequence. Deletion sequence bases are represented in lower case letters.

Junction	Nucleotide Sequence	Seq. ID
BCGΔ1a	CTGGTCGACGATTGGCACAT|gcagccgtgggtgccgccgg	1

BCGΔ1b	gtgtcttcatcggcttccac|CCAGCCGCCCGGATCCAGCA	2
BCGΔ2a	CAACTCCACGGCGACCACCC|gcgcccccgctcgcactaga	3
BCGΔ2b	gcccacccggtcgagcaccc|CGATGATCTTCTGTTTGACC	4
BCGΔ3a	CACCTCGACCACGGCCAACC|gtggacctgtgagatacact	5
BCGΔ3b	tcagcagtccacggccaacc|CCGCACCAACACCTTCCACC	6
BCGΔ1ab	CTGGTCGACGATTGGCACAT|CCAGCCGCCCGGATCCAGCA	7
BCGΔ2ab	CAACTCCACGGCGACCACCC|CGATGATCTTCTGTTTGACC	8
BCGΔ3ab	CACCTCGACCACGGCCAACC|CCGCACCAACACCTTCCACC	9

Where a deletion is detected by determining the presence or absence of sequences contained within the deletion (deletion sequences), the absence of deletion sequences indicates the presence of a deletion and thus an avirulent phenotype. Conversely, the presence of deletion sequences indicates the absence of a deletion. Deletion sequences that provide suitable targets for detecting the deletions of the present invention are provided in Figures 1, 2 and 3.

A) Isolation of DNA for Detection of Mycobacterium Genomic Deletions

In a preferred embodiment, DNA is obtained from mycobacteria. As used herein, the term "mycobacteria" refers to any bacteria of the family *Mycobacteriaceae* (order *Actinomycetales*) and includes, but is not limited to, *Mycobacterium tuberculosis*, *Mycobacterium avium complex*, *Mycobacterium kansasii*, *Mycobacterium scrofulaceum*, *Mycobacterium bovis and Mycobacterium leprae*. These species and groups and others are described in Baron, S., ed. *Medical Microbiology*, 3rd Ed. (1991) Churchill Livingstone, New York, which is incorporated herein by reference.

The identification of deletions using a DNA marker requires that the DNA sequence be accessible to the particular probes used or to the components of the amplification system if the DNA sequence is to be amplified. In general, this accessibility is ensured by isolating the nucleic acids from the sample.

A variety of techniques for extracting nucleic acids from biological samples are known in the art. For example, see those described by Sambrook *et al.*, *Molecular Cloning - A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New

York, (1985), by Han, *et al. Biochemistry*, 26: 1617-1625 (1987) and by Du, *et al. Bio/Technology*, 10: 176-181 (1992), which are incorporated herein by reference.

Alternatively, if the sample is readily disruptable, the nucleic acid need not be purified prior to amplification by the PCR technique, *i.e.*, if the sample is comprised of cells, particularly peripheral blood lymphocytes or monocytes, lysis and dispersion of the intracellular components may be accomplished merely by suspending the cells in hypotonic buffer or boiling them in a low concentration of alkali (*i.e.* 10 mM NaOH).

In a preferred embodiment, DNA is extracted from mycobacteria as described in Example 1.

B) Detection of Deletions Using Hybridization Probes

In one embodiment the avirulence deletions are detected by contacting DNA obtained from the mycobacterium with a probe that specifically binds an entire deletion junction region or a subsequence of that region and does not specifically bind to any other DNA sequences in the sample. Alternatively, a probe that specifically binds the entire deleted region or subsequence of that region and does not specifically bind to any other sequences in the sample is also suitable. While such probes may be proteins, oligonucleotide probes are preferred. Typically, the sequence of the oligonucleotide probe is chosen to be complementary to a select subsequence unique to the deletion junction or the deletion sequence, whose presence or absence is to be detected. Under stringent conditions the probe will hybridize with the select subsequence forming a stable duplex.

The probe is typically labeled. Detection of the label in association with the target DNA indicates either the presence or absence of the deletion. The probe may be used to detect the deletion junction or deletion sequences directly in a DNA sample without amplification of the deletion subsequences. In one embodiment, unamplified DNA sequences are probed using a Southern blot. The DNA of the sample is immobilized, on a solid substrate, typically a nitrocellulose filter or a nylon membrane. The substrate-bound DNA is then hybridized with the labeled probe under stringent conditions and non-specifically hybridized probe is washed away. Labeled probe detected in association with the immobilized mycobacterial sequences (*e.g.* bound to the substrate) indicates the presence of deletion sequences (*e.g.* BCGΔ1, BCGΔ2, or BCGΔ3) and therefore the absence of the deletion. Means for detecting specific DNA sequences are well known to those of skill in

the art. Protocols for Southern blots as well as other detection methods are provided in Maniatis, *et al. Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, NY (1982), which is incorporated herein by reference.

In another embodiment, the mycobacterial DNA subsequences are themselves labeled. They are then hybridized, under stringent conditions, with a probe immobilized on a solid substrate. Detection of the label in association with the immobilized probe indicates the presence or absence of the deletion.

In a preferred embodiment, the deletion junction sequences or subsequences or the deletion sequences or subsequences may be amplified by a variety of DNA amplification techniques (for example via cloning, polymerase chain reaction, ligase chain reaction, transcription amplification, *etc.*) prior to detection using a probe. Because the copy number of mycobacterial sequences bearing the virulence-attenuating deletions is low, the use of unamplified mycobacterial DNA results in an assay of low sensitivity. Amplification of mycobacterial DNA increases sensitivity of the assay by providing more copies of possible target subsequences. In addition, by using labeled primers in the amplification process, the mycobacterial DNA sequences are labeled as they are amplified.

C) Selection of Probes for Detection of the Deletion Junction Sequences or the Deletion Sequences

Full length sequences are provided for the deletions BCGΔ1, BCGΔ2, and BCGΔ3 in Figures 1, 2 and 3 respectively. Using these sequence listings, one of skill in the art may easily determine appropriate probes or primers for the detection of the presence or absence of the deletion junctions or the deletion sequences. Generally speaking, a probe will be selected that hybridizes to the target junction sequences or deletion sequences, but not to other mycobacterial nucleic acid sequences under stringent conditions. The design of hybridization probes is well known in the art. See, for example, Sambrook *et al.*, *Molecular Cloning - A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, (1989), which is incorporated herein by reference.

In a preferred embodiment, the probe is an oligonucleotide sequence complementary to a subsequence comprising a deletion junction (*e.g.* BCGΔ1a, BCGΔ1b, BCGΔ2a, BCGΔ2b, BCGΔ3a, BCGΔ3b, BCGΔ1ab, BCGΔ2ab, and BCGΔ3ab) or a

sequence complementary to a subsequence of a deletion sequence (*e.g.* BCGΔ1, BCGΔ2, and BCGΔ3). The probe preferably has destabilizing mismatches with subsequences from other regions of the mycobacterial genome.

The exact length of the probe depends on many factors including the length of conserved regions around the deletions, the degree of sequence specificity desired, and the amount of internal complementarity within the probe. Such probes are preferably 17 to 25 bases in length. One of skill will recognize that longer probes specifically hybridize at higher temperatures. Generally, stringent conditions are selected to be about 5°C to 20°C, more preferably about 10°C, lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength and pH. Under stringent conditions, the probe will specifically hybridize to a nucleic acid sequence from an avirulent mycobacterium such as BCG, but not to a nucleic acid sequence from a virulent mycobacterium such as MTB or MBV. Alternatively, Under stringent conditions, the probe will specifically hybridize to a nucleic acid sequence from a avirulent mycobacterium such as MTB or MBV, but not to a nucleic acid sequence from an avirulent mycobacterium such as BCG.

Oligonucleotide probes can be prepared by any suitable method, including, for example, cloning and restriction of appropriate sequences and direct chemical synthesis by a method such as the phosphotriester method of Narang *et al. Meth. Enzymol*, 68: 90-99 (1979); the phosphodiester method of Brown *et al.*, *Meth. Enzymol.* 68:109-151 (1979); the diethylphosphoramidite method of Beaucage *et al.*, *Tetra. Lett.*, 22: 1859-1862 (1981); and the solid support method of U.S. Patent No. 4,458,066.

Probe detectability may be increased by the attachment of a label. As used herein, a label is any composition detectable by spectroscopic, photochemical, biochemical, immunochemical, electrical, optical or chemical means. Useful labels in the present invention include magnetic beads (*e.g.* Dynabeads™), fluorescent dyes (*e.g.*, fluorescein isothiocyanate, texas red, rhodamine, and the like), radiolabels (*e.g.*, ^{3}H, ^{125}I, ^{35}S, ^{14}C, or ^{32}P), enzymes (*e.g.*, horse radish peroxidase, alkaline phosphatase and others commonly used in an ELISA), and colorimetric labels such as colloidal gold or colored glass or plastic (*e.g.* polystyrene, polypropylene, latex, *etc.*) beads.

Methods for attaching labels to probes, primers, and antibodies are well known to those of skill in the art. For example, the probe can be labeled at the 5'-end with ^{32}P by incubating the probe with ^{32}P-ATP and polynucleotide kinase (see Perbal, *A*

Practical Guide to Molecular Cloning, 2nd ed. John Wiley, N.Y. (1988)). Other labels may be joined to the probe directly or through linkers. They may be located at the ends of the probe or internally. Methods of attaching labels may be found in Connell, *et al.*, *Bio/Techniques* 5: 342 (1987), U.S. Patent Nos. 4,914,210, 4,391,904 and 4,962,029, which are incorporated herein by reference. In addition, kits for labelling oligonucleotides are widely available. See, for example, Boehringer Mannheim Biochemicals (Indianapolis, IN) for "Genius" labeling kits based on dioxigenin technology and Clonetech (South San Francisco, CA) for a variety of direct and indirect oligonucleotide labeling reagents.

D) Detection of Deletions Conferring Avirulence Through Amplification of Unique Subsequences

Deletions are particularly amenable to detection without the use of a hybridization probe. In a preferred embodiment, subsequences are amplified that include a deletion junction. The amplified deletion junction may be a "spanning" deletion junction in which case where the deletion is present (*i.e.* the deletion sequences are absent), the amplification product is a specific DNA incorporating the deletion junction sequence spanning the deletion (*e.g.* incorporating flanking sequences from both sides of the deleted sequence). Where the deletion is absent (*i.e.* deletion sequences are present) and primers are selected so that there are no priming sites within the deletion sequences, amplification is non-existent or alternatively provides a complex mixture of non-specifically amplified fragments. Alternatively, amplification primers may be selected that specifically hybridize to deletion sequences, as long as they are selected to amplify sequences that are distinguishable from the sequence amplified when the deletion is present.

Alternatively, the amplification product may be subsequence of a "terminal" deletion junction in which case absence of the deletion (*i.e.* the deletion sequences are present) will result in the amplification of the specifically targeted nucleic acid. Conversely, where the deletions are present (*i.e.* the deletion sequences are absent) there will be no specific amplification of a terminal deletion junction.

Amplification products may be separated by size for characterization. Size separation may be accomplished by a variety of means known to those of skill in the art.

These methods include, but are not limited to electrophoresis, density gradient centrifugation, liquid chromatography, and capillary electrophoresis. In a preferred embodiment, the fragments are separated by agarose gel electrophoresis. The bands are then stained with a marker to visualize them such as ethidium bromide and the gel is visualized, *e.g.*, using ultraviolet light.

As described above, an agarose gel typically shows 1 band if the deletion is present, reflecting amplification of the deletion-spanning sequence. Where the deletion is absent, amplification results in either no bands, where there are no sequences within the deletion to which the amplification primers may hybridize, or a smear where there is non-specific amplification, or a series of discrete bands distinguishable from the band representing the deletion-spanning sequence where primers are chosen that hybridize to deletion sequences.

E) Selection of Primers for Amplification of Avirulence Deletions

Amplification of deletion junction sequences or subsequences or deletion sequences or subsequences may be accomplished by methods well known in the art, which include, but are not limited to polymerase chain reaction (PCR) (Innis, *et al.*, *PCR Protocols. A guide to Methods and Application.* Academic Press, Inc. San Diego, (1990), which is incorporated herein by reference), ligase chain reaction (LCR) (see Wu and Wallace, *Genomics*, 4: 560 (1989), Landegren, *et al.*, *Science*, 241: 1077 (1988) and Barringer, *et al.*, *Gene*, 89: 117 (1990), which are incorporated herein by reference), transcription amplification (see Kwoh, *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.)*, 86: 1173 (1989) which is incorporated herein by reference), and self-sustained sequence replication (see Guatelli, *et al.*, *Proc. Nat. Acad. Sci. (U.S.A.)*, 87: 1874 (1990) which is incorporated herein by reference), each of which provides sufficient amplification so that the target sequence can be detected by nucleic acid hybridization to a probe or by electrophoretic separation. Alternatively, methods that amplify the hybridization probe to detectable levels can be used, such as Qβ-replicase amplification. See, for example, Kramer, *et al. Nature*, 339: 401 (1989), Lizardi, *et al. Bio/Technology*, 6: 1197 (1988), and Lomell, *et al.*, *Clin. Chem.* 35: 1826 (1989) which are incorporated herein by reference.

In a preferred embodiment, amplification is by polymerase chain reaction using a pair of primers that flank and thereby amplify a selected deletion junction subsequence. Selection of primers is readily apparent to one of skill in the art using the sequence listings of the present invention. For example, a pair of PCR primers 5'-TCGACGATTGGCACAT-3' (T_m=55°C) and 5'-TCCCTCCCTGTATTTGTAT-3' (T_m=56°C) will amplify a 469 base pair sequence including the BCGΔ1a deletion junction, while 5'-CGTTCTTCGGAGGTTTC-3' (T_m=56°C) and 5'-GGCGGCTGGGTGGA-3' (T_m=60°C) will amplify a 471 base pair sequence including the BCGΔ1b deletion junction.

F) Detection of Deletions through Detection of Expression Products of Deletion Sequences

In addition to the detection of deletions by the detection of either the deletion junction sequences or the deletion sequences, one may detect the absence of the deletion by detecting the expression products of the deletion sequences. Thus, for example, where the deletion sequences express a protein, the presence of that protein indicates the absence of the deletion and thus is indicative of a virulent (non BCG-like) phenotype. Such proteins are referred to herein as "deletion polypeptides".

Means of determining proteins expressed by particular nucleic acid sequences are well known to those of skill in the art. Typically this involves determining the longest open reading frame. This may be aided by the identification of initiation sites (*e.g.* ribozome binding sites). The protein encoded by the largest open reading frame is determined using codon preferences for the specific organism from which the nucleic acid is obtained. The polypeptide sequence listing may then be compared against a sequence database, *e.g.* GenBank, to determine other sequences sharing substantial sequence identity with the calculated sequence. The expression of the protein may be verified by isolating and then sequencing proteins having the predicted length and charge characteristics.

Once deletion polypeptides are identified they may be detected by routine methods well known to those of skill in the art. Typically this involves isolating and then detecting the polypeptide. The polypeptide may be isolated by a number of means well known to those of skill in the art. This includes typical methods of protein

purification such as high performance liquid chromatography (HPLC), electrophoresis, capillary electrophoresis, hyperdiffusion chromatography, thin layer chromatography, and the like. Methods of purifying and detecting proteins are well known to those of skill in the art (see, *e.g.*, *Methods in Enzymology Vol. 182: Guide to Protein Purification*, M. Deutscher, ed. Vol. 182 (1990), which is incorporated herein by reference).

Alternatively, deletion polypeptides sequences may be detected using immunoassays utilizing antibodies specific for the deletion polypeptides. The production of such antibodies and their use in immunoassays is detailed below.

G) Antibodies to Deletion Polypeptides

Antibodies can be raised to the polypeptides encoded by the nucleic acids corresponding to the open reading frames present in the deletion regions of the present invention (deletion polypeptides). As used herein "antibodies" include immunoglobulin or a population of immunoglobins which specifically bind to an antigen. Thus an antibody may be monoclonal or polyclonal including individual, allelic, strain, or species variants, and fragments thereof, both in their naturally occurring (full-length) forms and in recombinant forms. Additionally, antibodies can be raised to these polypeptides in either their native configurations or in non-native configurations. Anti-idiotypic antibodies may also be used.

1) Antibody Production

A number of immunogens may be used to produce antibodies specifically reactive with deletion polypeptides. Recombinant polypeptides are the preferred immunogen for the production of monoclonal or polyclonal antibodies. Naturally occurring polypeptides may also be used either in pure or impure form. Synthetic peptides made using sequences described herein may also used as immunogens for the production of antibodies.

Recombinant polypeptides are expressed in eukaryotic or prokaryotic cells and purified using standard techniques. The polypeptide is injected into an animal capable of producing antibodies. Either monoclonal or polyclonal antibodies may be generated for subsequent use in immunoassays to measure the presence and quantity of the polypeptide.

Methods of producing polyclonal antibodies are known to those of skill in the art. In brief, an immunogen, preferably a purified deletion polypeptide is mixed with an adjuvant and animals are immunized with the mixture. The animal's immune response to the immunogen preparation is monitored by taking test bleeds and determining the titer of reactivity to the polypeptide of interest. When appropriately high titers of antibody to the immunogen are obtained, blood is collected from the animal and antisera are prepared. Further fractionation of the antisera to enrich for antibodies reactive to the polypeptide is performed where desired. *See, e.g.*, Coligan (1991) *Current Protocols in Immunology* Wiley/Greene, NY; and Harlow and Lane (1989) *Antibodies: A Laboratory Manual* Cold Spring Harbor Press, NY, which are incorporated herein by reference.

Monoclonal antibodies may be obtained by various techniques familiar to those skilled in the art. Description of techniques for preparing such monoclonal antibodies may be found in, *e.g.*, Stites *et al.* (eds.) *Basic and Clinical Immunology* (4th ed.) Lange Medical Publications, Los Altos, CA, and references cited therein; Harlow and Lane (1988) *Antibodies: A Laboratory Manual* CSH Press; Goding (1986) *Monoclonal Antibodies: Principles and Practice* (2d ed.) Academic Press, New York, NY; and particularly in Kohler and Milstein (1975) *Nature* 256: 495-497, which discusses one method of generating monoclonal antibodies.

Summarized briefly, this method involves injecting an animal with an immunogen. The animal is then sacrificed and cells taken from its spleen, which are then fused with myeloma cells (*See*, Kohler and Milstein (1976) *Eur. J. Immunol.* 6: 511-519, incorporated herein by reference). The result is a hybrid cell or "hybridoma" that is capable of reproducing *in vitro*.

Colonies arising from single immortalized cells are screened for production of antibodies of the desired specificity and affinity for the antigen, and yield of the monoclonal antibodies produced by such cells is enhanced by various techniques, including injection into the peritoneal cavity of a vertebrate host. Alternatively, one may isolate DNA sequences which encode a monoclonal antibody or a binding fragment thereof by screening a DNA library from human B cells according to the general protocol outlined by Huse *et al.* (1989) *Science* 246: 1275-1281. In this manner, the individual antibody species obtained are the products of immortalized and cloned single B

cells from the immune animal generated in response to a specific site recognized on the immunogenic substance.

Other suitable techniques involve selection of libraries of antibodies in phage or similar vectors. See, Huse *et al. Science* 246: 1275-1281 (1989); and Ward, *et al. Nature* 341: 544-546 (1989). The polypeptides and antibodies of the present invention are used with or without modification, including chimeric antibodies. Frequently, the polypeptides and antibodies will be labeled by joining, either covalently or non-covalently, a substance which provides for a detectable signal. A wide variety of labels and conjugation techniques are known and are reported extensively in both the scientific and patent literature. Suitable labels include radionucleotides, enzymes, substrates, cofactors, inhibitors, fluorescent moieties, chemiluminescent moieties, magnetic particles, and the like. Patents, teaching the use of such labels include U.S. Patent Nos. 3,817,837; 3,850,752; 3,939,350; 3,996,345; 4,277,437; 4,275,149; and 4,366,241. Also, recombinant immunoglobulins may be produced. See, Cabilly, U.S. Patent No. 4,816,567; and Queen *et al. Proc. Nat'l Acad. Sci. USA* 86: 10029-10033 (1989).

Antibodies, including binding fragments and single chain versions, against predetermined fragments of deletion polypeptides can be raised by immunization of animals with conjugates of the fragments with carrier proteins as described above. Monoclonal antibodies are prepared from cells secreting the desired antibody. These antibodies can be screened for binding to normal or defective polypeptides, or screened for agonistic or antagonistic activity, *e.g.*, mediated through a receptor. These monoclonal antibodies will usually bind with at least a K_D of about 1 mM, more usually at least about 300 μM, and most preferably at least about 0.1 μM or better.

The antibodies of this invention can also be used for affinity chromatography in isolating deletion polypeptides. Columns can be prepared where the antibodies are linked to a solid support, *e.g.*, particles, such as agarose, Sephadex, or the like, where a bacterial lysate, or recombinant cell lysate is passed through the column, washed, and treated with increasing concentrations of a mild denaturant, whereby purified deletion polypeptides are released.

The antibodies can be used to screen expression libraries for particular expression products. Usually the antibodies in such a procedure will be labeled with a moiety allowing easy detection of presence of antigen by antibody binding.

In a preferred embodiment, antibodies to deletion polypeptides are used for the identification of cell populations expressing the polypeptides. By assaying the expression products of cells expressing the polypeptides it is possible to diagnose bacterial infections.

Antibodies raised against each polypeptide are useful to raise anti-idiotypic antibodies. These will be useful in detecting or diagnosing various immunological conditions related to the presence of the respective antigens.

2) Immunoassays

A particular deletion polypeptide can be measured by a variety of immunoassay methods. For a review of immunological and immunoassay procedures in general, see Stites and Terr (eds.) 1991 *Basic and Clinical Immunology* (7th ed.). Moreover, the immunoassays of the present invention can be performed in any of several configurations, *e.g.*, those reviewed in Maggio (ed.) (1980) *Enzyme Immunoassay* CRC Press, Boca Raton, Florida; Tijan (1985) "Practice and Theory of Enzyme Immunoassays," *Laboratory Techniques in Biochemistry and Molecular Biology*, Elsevier Science Publishers B.V., Amsterdam; and Harlow and Lane *Antibodies, A Laboratory Manual*, *supra*, each of which is incorporated herein by reference. See also Chan (ed.) (1987) *Immunoassay: A Practical Guide* Academic Press, Orlando, FL; Price and Newman (eds.) (1991) *Principles and Practice of Immunoassays* Stockton Press, NY; and Ngo (ed.) (1988) *Non-isotopic Immunoassays* Plenum Press, NY.

Immunoassays for measurement of deletion polypeptides can be performed by a variety of methods known to those skilled in the art. In brief, immunoassays to measure the protein can be, *e.g.*, competitive or noncompetitive binding assays. In competitive binding assays, the sample to be analyzed competes with a labeled analyte for specific binding sites on a capture agent bound to a solid surface. Preferably the capture agent is an antibody specifically reactive with a deletion polypeptide produced as described above. The concentration of labeled analyte bound to the capture agent is inversely proportional to the amount of free analyte present in the sample.

In a competitive binding immunoassay, the deletion polypeptide present in the sample competes with labelled protein for binding to a specific binding agent, for example, an antibody specifically reactive with a particular deletion polypeptide. The binding agent is, *e.g.*, bound to a solid surface to produce separation of bound labelled polypeptide from the unbound labelled polypeptide. Alternately, the competitive binding assay may be conducted in liquid phase and any of a variety of techniques known in the art may be used to separate the bound labelled protein from the unbound labelled protein. Following separation, the amount of bound labeled protein is determined. The amount of polypeptide present in the sample is inversely proportional to the amount of labelled polypeptide binding.

Alternatively, a homogenous immunoassay may be performed in which a separation step is not needed. In these immunoassays, the label on the protein is altered by the binding of the protein to its specific binding agent. This alteration in the labelled protein results in a decrease or increase in the signal emitted by label, so that measurement of the label at the end of the immunoassay allows for detection or quantitation of the polypeptide.

Deletion polypeptides may also be detected by a variety of noncompetitive immunoassay methods. For example, a two-site, solid phase sandwich immunoassay may be used. In this type of assay, a binding agent for the protein, for example an antibody, is attached to a solid support. A second protein binding agent, which is also an antibody, and which binds the protein at a different site, is labelled. After binding at both sites on the protein, the unbound labelled binding agent is removed and the labelled binding agent bound to the solid phase is measured. The amount of labelled binding agent bound is directly proportional to the amount of polypeptide in the sample.

Western blot analysis can be used to determine the presence of a deletion polypeptide in a sample. Electrophoresis is carried out, for example, on a bacterial sample suspected of containing the deletion polypeptide. Following electrophoresis to separate the proteins, and transfer of the proteins to a suitable solid support such as a nitrocellulose filter, the solid support is incubated with an antibody reactive with the protein. This antibody is labelled, or alternatively may be it is detected by subsequent incubation with a second labelled antibody that binds the primary antibody.

The immunoassay formats described above employ labelled assay components. The label can be in a variety of forms as described above. The choice of label depends on sensitivity required, ease of conjugation with the compound, stability requirements, and available instrumentation. For a review of various labelling or signal producing systems which may be used, see U.S. Patent No. 4,391,904, which is incorporated herein by reference.

Antibodies reactive with a particular protein can also be measured by a variety of immunoassay methods. For a review of immunological and immunoassay procedures applicable to the measurement of antibodies by immunoassay techniques, see Stites and Terr (eds.) *Basic and Clinical Immunology* (7th ed.) *supra*; Maggio (ed.) *Enzyme Immunoassay*, *supra*; and Harlow and Lane *Antibodies, A Laboratory Manual*, *supra*.

In brief, immunoassays to measure antisera reactive with polypeptides include competitive and noncompetitive binding assays. In competitive binding assays, the sample analyte competes with a labeled analyte for specific binding sites on a capture agent bound to a solid surface. Preferably the capture agent is a purified recombinant deletion polypeptide as described above. Other sources of polypeptides, including isolated or partially purified naturally occurring protein, can also be used. Noncompetitive assays are typically sandwich assays, in which the sample analyte is bound between two analyte-specific binding reagents. One of the binding agents is used as a capture agent and is bound to a solid surface. The second binding agent is labelled and is used to measure or detect the resultant complex by visual or instrument means. A number of combinations of capture agent and labelled binding agent can be used. A variety of different immunoassay formats, separation techniques and labels can be also be used similar to those described above for the measurement of deletion polypeptides.

II. Preparation of Deletion-Containing Mycobacteria

Mycobacteria containing specific deletions may be prepared by using methods of homologous recombination well known to those of skill in the art. In brief, homologous recombination is a natural cellular process which results in the scission of two nucleic acid molecules having identical or substantially similar (*i.e.* "homologous") sequences, and the ligation of the two molecules such that one region of each initially

present molecule is now ligated to a region of the other initially present molecule (Sedivy, *Bio/Technol.*, 6: 1192-1196 (1988).

Homologous recombination is exploited by a number of various methods of "gene targeting" well known to those of skill in the art. (*see, for example*, Mansour *et al. Nature*, 336: 348-352 (1988); Capecchi *Trends Genet.* 5: 70-76 (1989); Capecchi *Science* 244: 1288-1292 (1989); Capecchi *et al.* pages 45-52 In: *Current Communications in Molecular Biology*, Capecchi, M.R. (ed.), Cold Spring Harbor Press, Cold Spring Harbor, N.Y. (1989); Frohman *et al. Cell* 56: 145-147 (1989)). Some approaches focus on increasing the frequency of recombination between two DNA molecules by treating the introduced DNA with agents which stimulate recombination (*e.g.* trimethylpsoralen, UV light, *etc.*), however, most approaches utilize various combinations of selectable markers to facilitate isolation of the transformed cells.

One such selection method is termed positive/negative selection (PNS) (Thomas and Cappechi *Cell* 51: 503-512 (1987)). This method involves the use of two selectable markers: one a positive selection marker such as the bacterial gene for neomycin resistance (*neo*^{r}); the other a negative selection marker such as the herpes virus thymidine kinase (*tk*) gene. *Neo*^{r} confers resistance to the drug G-418, while herpes *tk* renders cells sensitive to the nucleoside analog gangcyclovir (GANC) or 1-(2-deoxy-2-fluoro-b-d-arabinofuranosyl)-5-iodouracil (FIAU). The DNA encoding the positive selection marker in the transgene (*e.g.* *neo*^{R}) is generally linked to an expression regulation sequence that allows for its independent transcription in mycobacteria. It is flanked by first and second sequence portions of at least a part of the deletion or deletion flanking sequences.

These first and second sequence portions target the transgene to a specific nucleotide sequence. A second independent expression unit capable of producing the expression product for a negative selection marker, *e.g.* for herpes virus tk is positioned adjacent to or in close proximity to the distal end of the first or second portions of the first DNA sequence. Upon transfection, some of the mycobacteria incorporate the transgene by random integration, others by homologous recombination between the endogenous allele and sequences in the transgene. As a result, one copy of the targeted nucleic acid is disrupted by homologous recombination with the-transgene with simultaneous loss of the sequence encoding herpes tk gene. Random integrants, which

occur via the ends of the transgene, contain herpes tk and remain sensitive to GANC or FIAU. Therefore, selection, either sequentially or simultaneously with G418 and GANC enriches for transfected mycobacteria containing the transgene integrated into the genome by homologous recombination.

Methods of homologous recombination in mycobacteria are described in greater detail by Ganjam *et al. Proc. Natl. Acad. Sci. USA*, 88: 5433-5437 (1991) and Aldovini *et al., J. Bacteriol.*, 175: 7282-7289 (1993) which are incorporated herein by reference.

III. Screening for Drug Susceptibility/Therapeutics

The expression products of the open reading frames in the BCGΔ1, BCGΔ2, and BCGΔ3 deletions of the present invention are targets for anti-mycobacterial drugs. To determine particularly suitable drug targets, open reading frames and surrounding expression control sequences are introduced into avirulent strains of mycobacteria, alone or in combination with other open reading frame regions to determine which regions are critical for virulence. Once particular genes are identified as critical for virulence, anti-mycobacterial agents are designed to inhibit expression of the critical genes, or to attack the critical gene products. For instance, antibodies are generated against the critical gene products and used as prophylactic or therapeutic agents. Alternatively, small molecules can be screened for the ability to selectively inhibit expression of the critical gene products, *e.g.*, using recombinant expression systems which include the gene's endogenous promoter. These small molecules are then used as therapeutics, or prophylactic agents to inhibit mycobacterial virulence.

In another embodiment, anti-mycobacterial agents which render a virulent mycobacterium avirulent can be operably linked to expression control sequences and used to transform a virulent mycobacterium. Such anti-mycobacterial agents inhibit the replication of a specified mycobacterium upon transcription or translation of the agent in the mycobacterium.

Such transformed mycobacteria are useful as vaccine components, and as components of immunological infectivity assays. For instance, an animal's blood can be monitored for the presence of anti-mycobacterial antibodies using the procedures described herein, using transformed avirulent mycobacterial components in various

immunological assays. Anti-mycobacterial agents useful in this invention include, without limitation, antisense genes, ribozymes, decoy genes, transdominant proteins and suicide genes.

An antisense nucleic acid is a nucleic acid that, upon expression, hybridizes to a particular mRNA molecule, to a transcriptional promoter or to the sense strand of a gene. By hybridizing, the antisense nucleic acid interferes with the transcription of a complementary DNA, the translation of an mRNA, or the function of a catalytic RNA. Antisense molecules useful in this invention include those that hybridize to gene transcripts in the region of the deletions of the invention, particularly deletion region 1.

A ribozyme is a catalytic RNA molecule that cleaves other RNA molecules having particular nucleic acid sequences. Ribozymes useful in this invention are those that cleave deletion gene transcripts. Examples include hairpin and hammerhead ribozymes.

A decoy nucleic acid is a nucleic acid having a sequence recognized by a regulatory DNA binding protein (*i.e.*, a transcription factor). Upon expression, the transcription factor binds to the decoy nucleic acid, rather than to its natural target in the genome. Useful decoy nucleic acid sequences include any sequence to which a transcription factor binds in the deletion regions of the present invention.

A transdominant protein is a protein whose phenotype, when supplied by transcomplementation, will overcome the effect of the native form of the protein. For instance, an avirulent mycobacterium can be rendered virulent by introducing transdominant proteins from deletion region 1.

A suicide gene produces a product which is cytotoxic. In the vectors of the present invention, a suicide gene is operably linked to an inducible expression control sequences which is stimulated upon infection of a cell by a mycobacterium.

IV. Use of Expressed "Deletion Proteins" in a Vaccine

The deletion polypeptides encoded by the open reading frames in BCGΔ1, BCGΔ2, and BCGΔ3 may be recombinantly expressed and used as components of immunological assays as described above or in vaccines. Expression of polypeptides

encoded by the open reading frames of the BCGΔ1, BCGΔ2, or BCGΔ3 deletions may be accomplished by means well known to those of skill in the art.

In brief, the expression of natural or synthetic nucleic acids encoding deletion polypeptides will typically be achieved by operably linking the DNA or cDNA to a promoter (which is either constitutive or inducible), followed by incorporation into an expression vector. The vectors can be suitable for replication and integration in either prokaryotes or eukaryotes. Typical expression vectors contain transcription and translation terminators, initiation sequences, and promoters useful for regulation of the expression of polynucleotide sequence encoding deletion polypeptides.

To obtain high level expression of a cloned gene, such as those polynucleotide sequences encoding deletion polypeptides, it is desirable to construct expression plasmids which contain, at the minimum, a promoter to direct transcription, a ribosome binding site for translational initiation, and a transcription/translation terminator. The expression vectors may also comprise generic expression cassettes containing at least one independent terminator sequence, sequences permitting replication of the plasmid in both eukaryotes and prokaryotes, *i.e.*, shuttle vectors, and selection markers for both prokaryotic and eukaryotic systems. For detailed techniques employed in the recombinant expression of deletion proteins *see*, for example, Sambrook, *et al.*, *Molecular Cloning: A Laboratory Manual* (2nd Ed., Vols. 1-3, Cold Spring Harbor Laboratory (1989)), *Methods in Enzymology*, Vol. 152: *Guide to Molecular Cloning Techniques* (Berger and Kimmel (eds.), San Diego: Academic Press, Inc. (1987)), or *Current Protocols in Molecular Biology*, (Ausubel, *et al.* (eds.), Greene Publishing and Wiley-Interscience, New York (1987), all of which are incorporated herein by reference.

The expressed deletion polypeptides may be used in a variety of assays. For example, the deletion polypeptides can be used as reagents in immunoblot assays to test whether a patient was previously exposed to virulent mycobacteria (*i.e.*, to test whether the patient has antibodies to the deletion polypeptide). These assays have the advantage of discriminating between previous exposure to an avirulent mycobacterium (*e.g.*, one used in a vaccine) and exposure to a virulent mycobacterium. Thus, vaccinated individuals can be tested for antibodies to the virulent mycobacterium without regard to whether the patient has been vaccinated with an avirulent mycobacterium.

The deletion polypeptides can also be used as antigenic vaccine components to direct antibodies to elements which are critical for virulence. These polypeptides can be added to existing vaccines (*e.g.*, those based upon avirulent mycobacteria and which lack the deletion polypeptide) to supplement the range of antigenicity conferred by the vaccine, or they may be used apart from other mycobacterial antigens. The vaccines of the invention contain as an active ingredient an immunogenically effective amount of a deletion polypeptide or of a recombinant vector which includes the deletion polypeptide. The immune response can include the generation of antibodies; activation of cytotoxic T lymphocytes (CTL) against cells presenting peptides derived from the polypeptides or other mechanisms well known in the art. See *e.g.* Paul *Fundamental Immunology Third Edition* published by Raven press New York (incorporated herein by reference) for a description of immune response. Useful carriers are well known in the art, and include, for example, thyroglobulin, albumins such as human serum albumin, tetanus toxoid, and polyamino acids such as poly(D-lysine:D-glutamic acid). The vaccines can also contain a physiologically tolerable (acceptable) diluent such as water, phosphate buffered saline, and further typically include an adjuvant. Adjuvants such as incomplete Freund's adjuvant, aluminum phosphate, aluminum hydroxide, or alum are materials well known in the art.

The compositions are suitable for single administrations or a series of administrations. When given as a series, inoculations subsequent to the initial administration are given to boost the immune response and are typically referred to as booster inoculations.

The vaccine compositions of the invention are intended for parenteral, topical, oral or local administration. Preferably, the pharmaceutical compositions are administered parenterally, *e.g.*, intravenously, subcutaneously, intradermally, or intramuscularly. Thus, the invention provides compositions for parenteral administration that comprise a solution of the agents described above dissolved or suspended in an acceptable carrier, preferably an aqueous carrier. A variety of aqueous carriers may be used, *e.g.*, water, buffered water, 0.4% saline, 0.3% glycine, hyaluronic acid and the like. These compositions may be sterilized by conventional, well known sterilization techniques, or may be sterile filtered. The resulting aqueous solutions may be packaged for use as is, or lyophilized, the lyophilized preparation being combined with a sterile

solution prior to administration. The compositions may contain pharmaceutically acceptable auxiliary substances as required to approximate physiological conditions, such as pH adjusting and buffering agents, tonicity adjusting agents, wetting agents and the like, for example, sodium acetate, sodium lactate, sodium chloride, potassium chloride, calcium chloride, sorbitan monolaurate, triethanolamine oleate, etc.

For solid compositions, conventional nontoxic solid carriers may be used which include, for example, pharmaceutical grades of mannitol, lactose, starch, magnesium stearate, sodium saccharin, talcum, cellulose, glucose, sucrose, magnesium carbonate, and the like. For oral administration, a pharmaceutically acceptable nontoxic composition is formed by incorporating any of the normally employed excipients, such as those carriers previously listed, and generally 10-95% of active ingredient and more preferably at a concentration of 25%-75%.

For aerosol administration, the polypeptides are preferably supplied in finely divided form along with a surfactant and propellant. The surfactant should be nontoxic, and preferably soluble in the propellant. Representative of such agents are the esters or partial esters of fatty acids containing from 6 to 22 carbon atoms, such as caproic, octanoic, lauric, palmitic, stearic, linoleic, linolenic, olesteric and oleic acids with an aliphatic polyhydric alcohol or its cyclic anhydride. Mixed esters, such as mixed or natural glycerides may be employed. A carrier can also be included, as desired, as with, *e.g.*, lecithin for intranasal delivery.

The amount of vaccine administered to the patient will vary depending upon the composition being administered, the physiological state of the patient and the manner of administration.

Live attenuated recombinant viruses which include the deletion polypeptide, such as recombinant vaccinia or adenovirus vectors, are convenient alternatives as vaccines because they are inexpensive to produce and are easily transported and administered. Vaccinia vectors and methods useful in immunization protocols are described, for example, in U.S. Patent No. 4,722,848, incorporated herein by reference.

Deletion sequences and subsequences of this invention may also be used in methods of genetic immunization. Briefly, genetic immunization involves transfecting

cells *in vivo* with nucleic acids encoding pathogen specific antigens. The transformed host cells then express the antigen thereby stimulating the host immune system.

In the present invention, antigen-encoding deletion region sequences are used to transform mammalian host cells thereby resulting in the expression of the antigen by the host. This provokes an immune response by the host against the expressed antigen thereby conferring immunity on the host. Methods of genetic immunization are well known to those of skill in the art (see, *e.g.*, Wang *et al. Proc. Natl. Acad. Sci. USA*, 90: 4156-4160 (1993); Ulmer *et al., Science*, 259: 1745-1749 (1993); Fynan *et al. DNA Cell Biol.*, 12: 785-789 (1993); Fynan *et al. Proc. Natl. Acad. Sci. USA*, 90: 11478-11482 (1993); Robinson *et al. Vaccine*, 11: 957-960 (1993); and Martinon *et al. Eur. J. Immunol.*, 23: 1719-1722 (1993), which are incorporated herein by reference.

VI. Use of Promoters within Deletion Sequences for Expression of Recombinant Proteins

Bacille Calmette-Guérin (BCG) contains all three deletions (BCGΔ1, BCGΔ2, and BCGΔ3) and yet is able to grow and reproduce indicating that the sequences contained within the deletion are not essential for bacterial viability. These deletion regions therefore make good target sites for the insertion of heterologous DNA as mycobacteria are tolerant of disruption of the native genome in these regions. The BCGΔ1, BCGΔ2, and BCGΔ3 deletion regions therefore provide suitable target sites for the incorporation of expression cassettes and the subsequent expression of exogenous gene products. The expression cassettes typically comprise a nucleic acid sequence under the control of a promoter. The promoter may be either constitutive or inducible. The cassette may additionally comprise a selectable marker such as an antibiotic resistance gene, a gene encoding a fluorescent marker (*e.g.* green fluorescent protein), or a gene encoding an enzymatic marker (*e.g.* ß-galactosidase).

Alternatively, genes under the control of endogenous promoters may be used as well. In one embodiment, reporter genes under the control of endogenous promoters found within the deletion sequences may be inserted at the deletion sites. These reporter genes may be utilized as an assay for antimycobacterial compounds that act by inhibiting transcription or translation of deletion sequences. Assaying for the

reporter gene product in the presence of an antimycobacterial compound provides a measure of efficacy of that compound in upregulating or downregulating deletion sequence genes. Methods of use of mycobacterial reporter gene assays to screen for drug activity are described by Cooksey *et al.*, *Antimicrob. Agents Chemother.*, 37: 1348-1352 (1993), and Jacobs *et al.*, *Science*, 260: 819-822 (1993) which are incorporated herein by reference.

EXAMPLES

The following examples are offered by way of illustration, not by way of limitation.

Example 1

Identification of Virulence-Attenuating Deletions

Bacterial Culture

All strains of Mycobacteria used in this study were maintained in 7H9 (Difco, Detroit Michigan, USA) media supplemented with OADC (BBL) or were grown on 7H11 agar supplemented with oleic acid albumin dextrose complex (OADC). *Escherichia coli* (strain DH5α or NM554) was used as a host for all recombinant plasmids and cosmids. *E. coli* was maintained in LB medium with or without agar. Carbenicillin (100 μg/ml) was used in place of ampicillin for the selection of all *E. coli* plasmids.

Extraction of High Molecular Weight DNA

High molecular weight chromosomal DNA was prepared by diluting a late log phase culture of the respective mycobacterium 1:10 into a liter of 7H9 medium containing 1.5% glycine and continuing growth for 4 to 5 days. The cells were then harvested by centrifugation, washed once in TE (pH 8.0) and resuspended in 4 ml of 25% sucrose in 10X TE. 100 μg of lysozyme was added and the preparation was incubated at 37°C for 2 hr followed by the addition of 100 μg of proteinase K and sarkosyl to a concentration of 1% weight/volume. Following overnight incubation at 65°C the mixture was extracted 4 times with chloroform isoamyl alcohol 24:1, once with phenol/chloroform (1:1), and twice again with chloroform isoamyl alcohol. The resulting high molecular weight DNA was then run on a CsCl gradient as described by

Hull *et al. Infect. Immun.*, 33: 933-938 (1981), which is incorporated herein by reference, and subsequently dialyzed against 4 changes of TE. BCG DNA was physically sheared by passage through a 22 gauge needle until an average size of 3-10 kb was obtained (20-25 passages). This DNA was then biotinylated using photobiotin (Clonetech, Palo Alto, California, USA) according to the method of Straus and Ausubel, *Proc. Natl. Acad. Sci. USA*, 87: 1889-1893 (1990), which is incorporated herein by reference.

DNA Subtraction

DNA subtraction was carried out between virulent *M. tuberculosis* H37Rv and avirulent BCG. H37R chromosomal DNA was selected because it was the most readily available chromosomal DNA from a virulent strain. In addition, *M. bovis* and *M tuberculosis* H37Rv are highly homologous.

M. bovis/M. tuberculosis specific probes were generated by the method of Straus and Ausubel, *supra.* with the following modifications. Sheared and biotinylated BCG DNA was used in a 10:1 excess for each round of subtraction. Wild type *M. tuberculosis* H37Rv DNA was digested with Sau3A to an average size of 1 kb. Hybridization conditions were 1M NaCl and 65 °C for 18 hours. Following five cycles (successive denaturation and reassociations) of subtraction, Sau3A1 adaptors (GACACTCTCGAGACATCACCGTCC and GATCGGACGGTGATGTCTCGAGAGTG were ligated to the subtraction product and amplified in a PCR reaction for 35 cycles (30 sec at 95°C, 30 sec at 55°C, and 3 min at 72°C). The *M. tuberculosis/M. bovis* specific probes were radiolabeled by using one strand of the adaptor (GACACTCTCGAGACATCACCGTCC) as a primer and labeling with ^{32}P dCTP using the Klenow fragment of DNA polymerase.

An *M. bovis* cosmid library was constructed in the BamH1 site of sCOS (Stratagene, La Jolla California, USA) with subsequent *in vitro* packaging and infection of *E. coli* strain NM554 (Stratagene). 600 colonies were picked to Nytran circular membranes and the membranes prepared according to the method of Grunstein and Hogness, *Proc. Natl. Acad. Sci. USA*, 72: 3961 (1975), which is incorporated herein by reference. These filters were then probed using the BCG subtracted probe and positive clones selected for further analysis. Cosmid DNA was prepared from selected clones by the method of Birnboim and Doly, *Nucleic Acids. Res.*, 7: 1513 (1973) which is

incorporated herein by reference. Restriction fragments that hybridize with the MTB/MBV specific probe were further subcloned into pGEM7z or pGEM5z (Promega, Madison, Wisconsin, USA) for deletion analysis.

Plasmid DNA for DNA sequencing was prepared using Qiagen minicolumns (Qiagen Inc. Chatsworth California, USA) and sequenced by the method of Henikoff, *Gene*, 28: 351-359 (1984), which is incorporated herein by reference, using the Erase A Base System (Promega). DNA sequencing reactions were run using a Perkin Elmer 9600 thermocycler and analyzed on an automated ABI sequencer. Analysis and assembly of contiguous DNA sequence was done using the ABI analysis software and SeQuencher sequence analysis software by Gene Clones Corp (Ann Arbor, Michigan, USA).

Deletion Region 1 (BCGΔ1)

Sequence analysis of over 16 kb of MBV region 1 and homologous regions in BCG revealed the precise junctions for the deletion in BCG. Eight open reading frames were identified with codon usage biases matching that of known MTB and MBV genes (see map Figure 4). The potential start and stop codons and predicted maximum protein coding capacity are listed in Figure 4. Consensus ribosomal binding site sequences were found near potential start codons for seven of eight open reading frames. TBLASTN and FASTA sequence homology analysis with each potential ORF-encoded protein revealed significant homologies for 3 of 8 open reading frames in region 1.

Most notable is the ORF1C homology to an unpublished and uncharacterized sequence listed in Genbank as *M. tuberculosis* antigen esat6. A 65 base pair repeated overlapping (repeated ~2 1/2 times) sequence was also recognized within the ORF1C (esat6) open reading frame. Also noteworthy are the significant homologies identified between ORF1H and bacterial serine proteases including *B. subtilus* subtilisin. Of the eight recognized open reading frames, four (ORFs 1B, 1C, 1D, and 1E) are located entirely within the 9 kb region deleted in BCG. One ORF traverses the BCG deletion junction in virulent *M. bovis*.

DNA probes from the 9 kb deletion in region 1 demonstrated that this region is absent in all BCG substrains and present in all virulent MBV and MTB strains tested. Furthermore, restriction fragment patterns observed in Southern blot analysis

with region 1 probes are non-polymorphic and identical in virulent MBV and MTB. This region has far fewer direct and indirect repeats than the regions 2 (BCGΔ2) and 3 (BCGΔ3) characterized below.

The sequence of a small region, estimated to be less than 20 bp between basepair coordinates 10654 and 10664 in region 1 has been recalcitrant to automated sequencing. Therefore, pending sequence confirmation, the base pair coordinates given in the region 1 map (Figure 4) are approximations. The precise sequence determination is likely to effect the Orf1E open reading frame.

Deletion Region 2 (BCGΔ2)

Sequence analysis of over 15 kb of MBV region 2 and homologous regions in BCG revealed the precise junctions for an 11 kb deletion in BCG. Thirteen open reading frames were identified with codon usage biases matching that of known MTB and MBV genes (see map Figure 5). The potential start and stop codons and predicted maximum protein coding capacity are also shown in Figure 5. Candidate consensus sequences resembling ribosomal binding sites were found near potential start codons for eight open reading frames. Of the thirteen open reading frames recognized in BCGΔ2, nine are located entirely within the 11 kb region deleted in most BCG strains while ORF2B2 and ORF2I traverse the deletion junctions.

TBLASTN and FASTA sequence homology analysis with each potential ORF-encoded protein revealed significant homologies for five open reading frames in BCGΔ2. A protein encoded by ORF2C exhibits striking similarity to the *E. coli* iciA protein which is thought to play a role in inhibiting and regulating the initiation of chromosomal replication. The iciA protein product is a member of the large LysR family of transcriptional regulatory proteins. Orf2F is highly homologous to an *S. typhimurium* ribonucleotide diphosphate reductase and a region of the *E. coli* and *S. typhimurium* proUVWX operon. Orf2H was found to have significant homology to *E. coli* and *S. typhimurium* permeases involved in aromatic amino acid transport and a eukaryotic cell retroviral receptor.

The Orf2G encoded protein was identical to the MTB mpt64 gene previously thought to encode a secreted antigen which is specifically expressed by MTB

and not BCG strains. Recent analysis of mpt64 expression revealed that three BCG substrains do express mpt64 (Moreau, Tokyo, Russian). Probes specific for mpt64 or other non-repetitive parts of region 2 hybridized to all MTB strains tested and the same three BCG substrains shown to express mpt64. Of interest is the finding that these three BCG substrains are derived from the original Pasteur strain prior to 1925. The current Pasteur strain and all strains derived from the original Pasteur strain after 1925, including the Connaught strain used in the subtractive analysis in this study, are deleted in the 11 kb DNA segment contained within BCGΔ2. These data indicate that an additional mutational event deleting the 11 kb segment of region 2, occurred in the BCG Pasteur strain sometime after 1925.

Southern blot analysis with probes from different segments of region 2 revealed a repetitive element located within a 2 kb segment (8-10 kb) of region 2. This repetitive element is ubiquitous in all tubercle bacilli tested. This element provides a marker suitable for RFLP analysis of mycobacterial strains.

Deletion Region 3 (BCGΔ3)

Sequence analysis of the almost 11 kb region 3 sequence and comparison to a homologous region in BCG precisely identified the deletion junctions for BCG. Twelve potential open reading frames were recognized in the region 3 sequence, seven of which are entirely located within the 9 kb region deleted in BCG. At least 9 ORFs in BCGΔ3 exhibit codon usage preferences comparable to that of the tubercle bacilli. Sequence homology analysis of presumptive protein sequences encoded by six open reading frames in region 3 revealed highly significant homology to listed sequences. Orfs3B, 3D, and 3E exhibit homology to phage sequences, suggesting a phage derivation for 4 or more kb of DNA in region 3. Homology to putative open reading frames in two *M. leprae* cosmids was also observed including homology to a putative *bid* gene encoding a protein involved in biotin synthesis. Also of interest was homology between ORF3A and an MTB sequence (*mce*) associated with cell invasion and intracellular survival.

Southern blot analysis with segments of region 3 deleted in BCG revealed that prototype lab strains of virulent MBV and MTB all carry deletion region 3 DNA. However, clinical isolates from PHRI are highly polymorphic or deleted in region 3.

This region contains many large direct and indirect repeats and, as mentioned above, at least 2 ORFs are homologous to phage sequences including homology to DNA invertases or recombinases. The repetitive nature of this region and the possible presence of a DNA recombinase could explain the polymorphisms observed in this region.

The sequence of a small region, estimated to be much less than 200 bp and located close to 9400 bp in Figure 3, was recalcitrant to automated sequencing and remains to be determined. Therefore, the base pair coordinates given in the region 3 map (Figure 6) 3' to the 9kb marker are approximations. The precise sequence determination of region is likely to effect the length of open reading frames 3H and 3L.

The foregoing subtractive analysis identified 3 regions in virulent *M. bovis* and *M. tuberculosis* prototype strains which are deleted in the avirulent BCG strain. The deletion located in region 2 may not have arisen in the original BCG Pasteur strain as this region is only deleted in strains derived from the original Pasteur strain after 1925. Region 3 is present in virulent MTB and MBV lab prototype strains (H37Rv, Erdman) and is highly polymorphic and at least partially deleted in the majority of MTB clinical isolates tested. Region 1 is apparently conserved and intact in all virulent MBV and MTB strains tested to date while all avirulent BCG strains tested to date are missing approximately 9kb from region 1.

Example 2

Screening and Identification of an Avirulent Mycobacterium

The ^{32} P labeled subtraction probe obtained in Example 1, was used to probe EcoRI and BamHI restricted chromosomal DNAs from BCG Connaught, *Mycobacterium bovis*, and various strains of *Mycobacterium tuberculosis* in a Southern blot. The hybridization was performed at 70°C in 6X SSC overnight.

The resulting Southern blot is illustrated in Figure 8. The probe showed no labeling of BCG reflecting the presence of all three deletions, while the other strains were labeled.

The above examples are provided to illustrate the invention but not to limit its scope. Other variants of the invention will be readily apparent to one of ordinary skill in the art and are encompassed by the appended claims. All publications, patents, and patent applications cited herein are hereby incorporated by reference.

WHAT IS CLAIMED IS:

1. A marker for an avirulent mycobacterium, said marker comprising a first nucleic acid that specifically hybridizes under stringent conditions with a second nucleic acid or a complement of said second nucleic acid where said second nucleic acid or complement of said second nucleic acid is selected from the group consisting of BCGΔ1a, BCGΔ1b, BCGΔ2a, BCGΔ2b, BCGΔ3a, BCGΔ3b, BCGΔ1ab, BCGΔ2ab, BCGΔ3ab, BCGΔ1, BCGΔ2, and BCGΔ3.

2. The marker of claim 1, wherein said marker specifically hybridizes under stringent conditions to a nucleic acid from BCG, but not to a nucleic acid from *Mycobacterium tuberculosis* or *Mycobacterium bovis*, or where said marker specifically hybridizes under stringent conditions to a nucleic acid from *Mycobacterium tuberculosis* or *Mycobacterium bovis*, but not to a nucleic acid from BCG.

3. The marker of claim 2, wherein said marker comprises a subsequence of a nucleic acid where said nucleic acid is selected from the group consisting of BCGΔ1a, BCGΔ1b, BCGΔ2a, BCGΔ2b, BCGΔ3a, BCGΔ3b, BCGΔ1ab, BCGΔ2ab, BCGΔ3ab, BCGΔ1, BCGΔ2, and BCGΔ3.

4. The marker of claim 2, wherein said marker is selected from the group consisting of BCGΔ1a, BCGΔ1b, BCGΔ2a, BCGΔ2b, BCGΔ3a, BCGΔ3b, BCGΔ1ab, BCGΔ2ab, BCGΔ3ab, BCGΔ1, BCGΔ2, and BCGΔ3.

5. The marker of claim 2, wherein said marker comprises a nucleic acid having at least 90 percent sequence identity with a sequence selected from the group consisting of BCGΔ1a, BCGΔ1b, BCGΔ2a, BCGΔ2b, BCGΔ3a, BCGΔ3b, BCGΔ1ab, BCGΔ2ab, BCGΔ3ab, BCGΔ1, BCGΔ2, and BCGΔ3.

6. The marker of claim 2, wherein said marker comprises a radioactive nucleotide probe.

7. The marker of claim 2, wherein said subsequence is a sequence selected from an open reading frame of a deletion, said deletion being selected from the group consisting of BCGΔ1, BCGΔ2, BCGΔ3.

8. A polypeptide encoded by a subsequence of a deletion sequence selected from the group consisting of BCGΔ1, BCGΔ2, and BCGΔ3.

9. The polypeptide of claim 8, wherein the subsequence is selected from an open reading frame (ORF) of a deletion, said deletion being selected from the group consisting of BCGΔ1, BCGΔ2, BCGΔ3.

10. An antibody that binds specifically to the polypeptide of claim 8.

11. A recombinant cell comprising a first nucleic acid that hybridizes under stringent conditions with a second nucleic acid or a complement of said second nucleic acid where said second nucleic acid or complement of said second nucleic acid is selected from the group consisting of BCGΔ1a, BCGΔ1b, BCGΔ2a, BCGΔ2b, BCGΔ3a, BCGΔ3b, BCGΔ1ab, BCGΔ2ab, BCGΔ3ab, BCGΔ1, BCGΔ2, and BCGΔ3.

12. The recombinant cell of claim 11, wherein the cell is a Mycobacterium.

13. The cell of claim 11, wherein the cell expresses a polypeptide encoded by an intact open reading frame from BCGΔ1, BCGΔ2, and BCGΔ3.

14. The cell of claim 11, wherein said cell is a mycobacterium having one or more deletions in the genomic regions selected from the group consisting of BCGΔ1, BCGΔ2, and BCGΔ3, wherein said deletions result in the attenuation of an otherwise virulent strain of mycobacterium and wherein said deletions are present in up to two of said regions.

15. The mycobacterium of claim 14, wherein said deletions comprise a deletion selected from the group consisting of BCGΔ1, BCGΔ2, and BCGΔ3.

16. A method of distinguishing between an attenuated and a virulent mycobacterium, said method comprising detecting the presence or absence of a first nucleic acid that hybridizes under stringent conditions with a second nucleic acid or a complement of said second nucleic acid where said second nucleic acid or complement of said second nucleic acid is selected from the group consisting of BCGΔ1a, BCGΔ1b, BCGΔ2a, BCGΔ2b, BCGΔ3a, BCGΔ3b, BCGΔ1ab, BCGΔ2ab, BCGΔ3ab, BCGΔ1, BCGΔ2, and BCGΔ3.

17. The method of claim 16, wherein said first nucleic acid specifically hybridizes under stringent conditions to a nucleic acid from BCG, but not to a nucleic acid from *Mycobacterium tuberculosis* or *Mycobacterium bovis*, or where said first nucleic acid specifically hybridizes under stringent conditions to a nucleic acid from *Mycobacterium tuberculosis* or *Mycobacterium bovis*, but not to a nucleic acid from BCG.

18. The method of claim 17, wherein said first sequence is amplified prior to detection.

19. The method of claim 17, wherein said first sequence is amplified by the polymerase chain reaction.

20. A method of claim 17, wherein said detecting comprises a Southern blot.

21. A method of claim 17, wherein said detecting comprises detecting a polypeptide encoded by said first nucleic acid.

22. The method of claim 21, wherein the polypeptide is encoded by an intact open reading frame of a nucleotide sequence selected from the group consisting of BCGΔ1, BCGΔ2, and BCGΔ3.

23. The method of claim 21, wherein the polypeptide is visualized by antibody hybridization.

24. A method for determining whether an attenuated or a virulent Mycobacterium is present in a sample comprising:

providing a first nucleic acid that hybridizes under stringent conditions with a second nucleic acid or a complement of said second nucleic acid where said second nucleic acid or complement of said second nucleic acid is selected from the group consisting of BCGΔ1a, BCGΔ1b, BCGΔ2a, BCGΔ2b, BCGΔ3a, BCGΔ3b, BCGΔ1ab, BCGΔ2ab, BCGΔ3ab, BCGΔ1, BCGΔ2, and BCGΔ3; and

hybridizing said first nucleic acid to the biological sample.

25. The method of claim 24, wherein said first nucleic acid specifically hybridizes under stringent conditions to a nucleic acid from BCG, but not to a nucleic acid from *Mycobacterium tuberculosis* or *Mycobacterium bovis*, or where said first nucleic acid specifically hybridizes under stringent conditions to a nucleic acid from *Mycobacterium tuberculosis* or *Mycobacterium bovis*, but not to a nucleic acid from BCG.

26. A method of producing an attenuated Mycobacterium species, said method comprising deleting from the genomic DNA of a virulent mycobacterium a first nucleic acid that specifically hybridizes under stringent conditions with a second nucleic acid or a complement of said second nucleic acid where said second nucleic acid or complement of said second nucleic acid is selected from the group consisting of BCGΔ1, BCGΔ2, and BCGΔ3.

Figure 1

WO 96/25519

PCT/US96/01938

```
      |  10      |  20      |  30      |  40      |  50      |  60      |  70      |  80      |  90      | 100
    1 GAATTCCTGC GCACCCTGAT CCTGTCGCTG GTGGCAATGA CTCATCCAGA TCAGGTGAAT CTCCTGCTCA CCGACTTCAA AGGTGGTTCA ACCTTCCTGG 100
  101 GAATGGAAAA GCTTCCGCAC ACTGCCGCTG TCGTCACCAA CATGGCCGAG GAAGCCGAGC TCGTCAGCCG GATGGGCGAG GTGTTGACCG GAGAACTCGA 200
  201 TCGGCGCCAG TCGATCCTCC GACAGGCCGG GATGAAAGTC GGCGCGGCCG GAGCCCTGTC CGGCGTGGCC GAATACGAGA AGTACCGCGA ACGCGGTGCC 300
  301 GACCTACCCC CGCTGCCAAC GCTTTTCGTC GTCGTCGACG AGTTCGCCGA GCTGTTGCAG AGTCACCCGG ACTTCATCGG GCTGTTCGAC CGGATCTGCC 400
  401 GCGTCGGGCG GTCGCTGAGG GTCCATCTGC TGCTGGCTAC CCAGTCGCTG CAGACCGGCG GTGTTCGCAT CGACAAACTG GAGCCAAACC TGACATATCG 500
  501 AATCGCATTG CGCACCACCA GCTCTCATGA ATCCAAGGCG GTAATCGGCA CACCGGAGGC GCAGTACATC ACCAACAAGG AGAGCGGTGT CGGGTTTCTC 600
  601 CGGGTCGGCA TGGAAGACCC GGTCAAGTTC AGCACCTTCT ACATCAGTGG GCCATACATG CCGCCGGCGG CAGGCGTCGA AACCAATGGT GAAGCCGTAG 700
  701 GGCCCGGTCA ACAGACCACT AGACAAGCCG CGCGCATTCA CAGGTTCACC GCGGCACCGG TTCTCGAGGA GGCGCCGACA CCGTGACCCG CGCCGGCGAC 800
  801 GATGCAAAGC GCAGCGATGA GGAGGAGCGG CGCCAACGGC CCGCGCCGGC GACGATGCAA AGCGCAGCGA TGAGGAGGAG CGGCGCGCAT GACTGCTGAA 900
  901 CCGGAAGTAC GGACGCTGCG CGAGGTTGTG CTGGACCAGC TCGGCACTGC TGAATCGCGT GCGTACAAGA TGTGGCTGCC GCCGTTGACC AATCCGGTCC 1000
 1001 CGCTCAACGA GCTCATCGCC CGTGATCGGC GACAACCCCT GCGATTTGCC CTGGGGATCA TGGATGAACC GCGCCGCCAT CTACAGGATG TGTGGGCCGT 1100
 1101 AGACGTTTCC GGGGCCGGCG GCAACATCGG TATTGGGGGC GCACCTCAAA CCGGGAAGTC GACGCTACTG CAGACGATGG TGATGTCGGC CGCCGCCACA 1200
 1201 CACTCACCGC GCAACGTTCA GTTCTATTGC ATCGACCTAG GTGGCGGCGG GCTGATCTAT CTCGAAAACC TTCCACACGT CGGTGGGGTA GCCAATCGGT 1300
 1301 CCGAGCCCGA CAAGGTCAAC CGGGTGGTCG CAGAGATGCA AGCCGTCATG CGGCAACGGG AAACCACCTT CAAGGAACAC CGAGTGGGCT CGATCGGGAT 1400
 1401 GTACCGGCAG CTGCGTGACG ATCCAAGTCA ACCCGTTGCG TCCGATCCAT ACGGCGACGT CTTTCTGATC ATCGACGGAT GGCCCGGTTT TGTCGGCGAG 1500
 1501 TTCCCCGACC TTGAGGGGCA GGTTCAAGAT CTGGCCGCCC AGGGGCTGGG GTTCGGCGTC CACGTCATCA TCTCCACGCC ACGCTGGACA GAGCTGAAGT 1600
 1601 CGCGTGTTCG CGACTACCTC GGCACCAAGA TCGAGTTCCG GCTTGGTGAC GTCAATGAAA CCCAGATCGA CCGGATTACC CGCGAGATCC CGGCGAATCG 1700
 1701 TCCGGGTCGG GCAGTGTCGA TGGAAAAGCA CCATCTGATG ATCGGCGTGC CCAGGTTCGA CGGCGTGCAC AGCGCCGATA ACCTGGTGGA GGCGATCACC 1800
 1801 GCGGGGGTGA CGCAGATCGC TTCCCAGCAC ACCGAACAGG CACCTCCGGT GCGGGTCCTG CCGGAGCGTA TCCACCTGCA CGAACTCGAC CCGAACCCGC 1900
 1901 CGGGACCAGA GTCCGACTAC CGCACTCGCT GGGAGATTCC GATCGGCTTG CGCGAGACGG ACCTGACGCC GGCTCACTGC CACATGCACA CGAACCCGCA 2000
 2001 CCTACTGATC TTCGGTGCGG CCAAATCGGC CAAGACGACC ATTGCCCACG CGATCGCGCG CGCCATTTGT GCCCGAAACA GTCCCCAGCA GGTGCGGTTC 2100
 2101 ATGCTCGCGG ACTACCGCTC GGGCCTGCTG GACGCGGTGC CGGACACCCA TCTGCTGGGC GCCGGCGCGA TCAACCGCAA CAGCGCGTCG CTAGACGAGG 2200
 2201 CCGCTCAAGC ACTGGCGGTC AACCTGAAGA AGCGGTTGCC GCCGACCGAC CTGACGACGG CGCAGCTACG CTCGCGTTCG TGGTGGAGCG GATTTGACGT 2300
 2301 CGTGCTTCTG GTCGACGATT GGCACATGCA GCCGTGGGTG CCGCCGGGGG GATGCCGCCG ATGGCACCGC TGGCCCCGTT ATTGCCGGCG GCGGCAGATA 2400
 2401 TCGGGTTGCA CATCATTGTC ACCTGTCAGA TGAGCCAGGC TTACAAGGCA ACCATGGACA AGTTCGTCGG CGCCGCATTC GGGTCGGGCG CTCCGACAAT 2500
 2501 GTTCCTTTCG GGCGAGAAGC AGGAATTCCC ATCCAGTGAG TTCAAGGTCA AGCGGCGCCC CCCTGGCCAG GCATTTCTCG TCTCGCCAGA CGGCAAAGAG 2600
 2601 GTCATCCAGG CCCCCTACAT CGAGCCTCCA GAAGAAGTGT TCGCAGCACC CCCAAGCGCC GGTTAAGATT ATTTCATTGC CGGTGTAGCA GGACCCGAGC 2700
 2701 TCAGCCCGGT AATCGAGTTC GGGCAATGCT GACCATCGGG TTTGTTTCCG GCTATAACCG AACGGTTTGT GTACGGGATA CAAATACAGG GAGGGAAGAA 2800
 2801 GTAGGCAAAT GGAAAAAATG TCACATGATC CGATCGCTGC CGACATTGGC ACGCAAGTGA GCGACAACGC TCTGCACGGC GTGACGGCCG GCTCGACGGC 2900
 2901 GCTGACGTCG GTGACCGGGC TGGTTCCCGC GGGGGCCGAT GAGGTCTCCG CCCAAGCGGC GACGGCGTTC ACATCGGAGG GCATCCAATT GCTGGCTTCC 3000
 3001 AATGCATCGG CCCAAGACCA GCTCCACCGT GCGGGCGAAG CGGTCCAGGA CGTCGCCCGC ACCTATTCGC AAATCGACGA CGGCGCCGCC GGCGTCTTCG 3100
 3101 CCTAATAGGC CCCCAACACA TCGGAGGGAG TGATCACCAT GCTGTGGCAC GCAATGCCAC CGGAGCTAAA TACCGCACGG CTGATGGCCG GCGCGGGTCC 3200
 3201 GGCTCCAATG CTTGCGGCGG CCGCGGGATG GCAGACGCTT TCGGCGGCTC TGGACGCTCA GGCCGTCGAG TTGACCGCGC GCCTGAACTC TCTGGGAGAA 3300
 3301 GCCTGGACTG GAGGTGGCAG CGACAAGGCG CTTGCGGCTG CAACGCCGAT GGTGGTCTGG CTACAAACCG CGTCAACACA GGCCAAGACC CGTGCGATGC 3400
 3401 AGGCGACGGC GCAAGCCGCG GCATACACCC AGGCCATGGC CACGACGCCG TCGCTGCCGG AGATCGCCGC CAACCACATC ACCCAGGCCG TCCTTACGGC 3500
 3501 CACCAACTTC TTCGGTATCA ACACGATCCC GATCGCGTTG ACCGAGATGG ATTATTTCAT CCGTATGTGG AACCAGGCAG CCCTGGCAAT GGAGGTCTAC 3600
 3601 CAGGCCGAGA CCGCGGTTAA CACGCTTTTC GAGAAGCTCG AGCCGATGGC GTCGATCCTT GATCCCGGCG CGAGCCAGAG CACGACGAAC CCGATCTTCG 3700
 3701 GAATGCCCTC CCCTGGCAGC TCAACACCGG TTGGCCAGTT GCCGCCGGCG GCTACCCAGA CCCTCGGCCA ACTGGGTGAG ATGAGCGGCC CGATGCAGCA 3800
 3801 GCTGACCCAG CCGCTGCAGC AGGTGACGTC GTTGTTCAGC CAGGTGGGCG GCACCGGCGG CGGCAACCCA GCCGACGAGG AAGCCGCGCA GATGGGCCTG 3900
 3901 CTCGGCACCA GTCCGCTGTC GAACCATCCG CTGGCTGGTG GATCAGGCCC CAGCGCGGGC GCGGGCCTGC TGCGCGCGGA GTCGCTACCT GGCGCAGGTG 4000
 4001 GGTCGTTGAC CCGCACGCCG CTGATGTCTC AGCTGATCGA AAAGCCGGTT TGCCCCCTCG GTGATGCCGG CGGCTGCTGC CGGATCGTCG GCGACGGGTG 4100
 4101 GCGCCGCTCC GGTGGGTGCG GGAGCGATGG GCCAGGGTGC GCAATCCGGC GGCTCCACCA GGCCGGGTCT GGTCGCGCCG GCACCGCTCG CGCAGGAGCG 4200
 4201 TGAAGAAGAC GACGAGGACG ACTGGGACGA AGAGGACGAC TGGTGAGCTC CCGTAATGAC AACAGACTTC CCGGCCACCC GGGCCGGAAG ACTTGCCAAC 4300
 4301 ATTTTGGCGA GGAAGGTAAA GAGAGAAAGT AGTCCAGCAT GGCAGAGATG AAGACCGATG CCGCTACCCT CGCGCAGGAG GCAGGTAATT TCGAGCGGAT 4400
 4401 CTCCGGCGAC CTGAAAACCC AGATCGACCA GGTGGAGTCG ACGGCAGGTT CGTTGCAGGG CCAGTGGCGC GGCGCGGCGG GGACGGCCGC CCAGGCCGCG 4500
```

Figure 1 continued, page 2 of 4

```
4501 GTGGTGCGCT TCCAAGAAGC AGCCAATAAG CAGAAGCAGG AACTCGACGA GATCTCGACG AATATTCGTC AGGCCGGCGT CCAATACTCG AGGGCCGACG 4600
4601 AGGAGCAGCA GCAGGCGCTG TCCTCGCAAA TGGGCTTCTG ACCCGCTAAT ACGAAAAGAA ACGGAGCAAA AACATGACAG AGCAGCAGTG GAATTTCGCG 4700
4701 GGTATCGAGG CCGCGGCAAG CGCAATCCAG GGAAATGTCA CGTCCATTCA TTCCCTCCTT GACGAGGGGA AGCAGTCCCT GACCAAGCTC GCAGCGGCCT 4800
4801 GGGGCGGTAG CGGTTCGGAG GCGTACCAGG GTGTCCAGCA AAAATGGGAC GCCACGGCTA CCGAGCTGAA CAACGCGCTG CAGAACCTGG CGCGGACGAT 4900
4901 CAGCGAAGCC GGTCAGGCAA TGGCTTCGAC CGAAGGCAAC GTCACTGGGA TGTTCGCATA GGGCAACGCC GAGTTCGCGT AGAATAGCGA AACACGGGAT 5000
5001 CGGGCGAGTT CGACCTTCCG TCGGTCTCGC CCTTTCTCGT GTTTATACGT TTGAGCGCAC TCTGAGAGGT TGTCATGGCG GCCGACTACG ACAAGCTCTT 5100
5101 CCGGCCGCAC GAAGGTATGG AAGCTCCGGA CGATATGGCA GCGCAGCCGT TCTTCGACCC CAGTGCTTCG TTTCCGCCGG CGCCCGCATC GGCAAACCTA 5200
5201 CCGAAGCCCA ACGGCCAGAC TCCGCCCCCG ACGTCCGACG ACCTGTCGGA GCGGTTCGTG TCGGCCCCGC CGCCGCCACC CCCACCCCCA CCTCCGCCTC 5300
5301 CGCCAACTCC GATGCCGATC GCCGCAGGAG AGCCGCCCTC GCCGGAACCG GCCGCATCTA AACCACCCAC ACCCCCCATG CCCATCGCCG GACCCGAACC 5400
5401 GGCCCCACCC AAACCACCCA CACCCCCCAT GCCCATCGCC GGACCCGAAC CGGCCCCACC CAAACCACCC ACACCTCCGA TGCCCATCGC CGGACCTGCA 5500
5501 CCCACCCCAA CCGAATCCCA GTTGCCGCCC CCCAGACCAC CGACACCACA AACGCCAACC GGAGCGCCGC AGCAACCGGA ATCACCGGCG CCCCACGTAC 5600
5601 CCTCGCACGG GCCACATCAA CCCCGGCGCA CCGCACCAGC ACCGCCCTGG GCAAAGATGC CAATCGGCGA ACCCCCGCCC GCTCCGTCCA GACCGTCTGC 5700
5701 GTCCCCGGCC GAACCACCGA CCCGCCCTGC CCCCCAACAC TCCCGACGTG CGCGCCGGGG TCACCGCTAT CGCACAGACA CCGAACGAAA CGTCGGGAAG 5800
5801 GTAGCAACTG GTCCATCCAT CCAGGCGCGG CTGCGGGCAG AGGAAGCATC CGGCGCGCAG CTCGCCCCCG GAACGGAGCC CTCGCCAGCG CCGTTGGGCC 5900
5901 AACCGAGATC GTATCTGGCT CCGCCCACCC GCCCCGCGCC GACAGAACCT CCCCCCAGCC CCTCGCCGCA GCGCAACTCC GGTCGGCGTG CCGAGCGACG 6000
6001 CGTCCGACCC CGATTTAGCC GCCCAACATG CCGCGGCGCA ACCTGATTCA ATTACGGCCG CAACCCACTG GCGGTCGTCG CCGCAAGCGT GCAGCGCCGG 6100
6101 GATGCTCGAC GCGACACAAG AAATCCTTAA GGCCGGCGGC CAAGGGGCCG AAGGTGAAGA AGGTGAAGCC CCAGAAACCG AAGGCCACGA AGCCGCCCAA 6200
6201 AGTGGTGTCG CAGCGCGGCT GGCGACATTG GGTGCATGCG TTGACGCGAA TCAACCTGGG CCTGTCACCC GACGAGAAGT ACGAGCTGGA CCTGCACGCT 6300
6301 CGAGTCCGCC GCAATCCCCG CGGGTCGTAT CAGATCGCCG TCGTCGGTCT CAAAGGTGGG GCTGGCAAAA CCACGCTGAC AGCAGCGTTG GGGTCGACGT 6400
6401 TGGCTCAGGT GCGGGCCGAC CGGATCCTGG CTCTAGACGC GGATCCAGGC GCCGGAAACC TCGCCGATCG GGTAGGCGCA CAATCGGGCG CGACCATCGC 6500
6501 TGATGTGCTT GCAGAAAAAG AGCTGTCGCA CTACAACGAC ATCCGCGCAC ACACTAGCGT CAATGCGGTC AATCTGGAAG TGCTGCCGGC ACCGGAATAC 6600
6601 AGCTCGGCGC AGCGCGCGCT CAGCGACGCC GACTGGCATT TCATCGCCGA TCCTGCGTCG AGGTTTTACA ACCTCGTCTT GGCTGATTGT GGGGCCGGCT 6700
6701 TCTTCGACCC GCTGACCCGC GGCGTGCTGT CCACGGTGTC CGGTGTCGTG GTCGTGGCAA GTGTCTCAAT CGACGGCGCA CAACAGGCGT CGGTCGCGTT 6800
6801 GGACTGGTTG CGCAACAACG GTTACCAAGA TTTGGCGAGC CGCGCATGCG TGGTCATCAA TCACATCATG CCGGGAGAAC CCAATGTCGC AGTTAAAGAC 6900
6901 CTGGTGCGGC ATTTCGAACA GCAAGTTCAA CCCGGCCGGG TCGTGGTCAT GCCGTGGGAC AGGCACATTG CGGCCGGAAC CGAGATTTCA CTCGACTTGC 7000
7001 TCGACCCTAT CTACAAGCGC AAGGTCCTCG AATTGGCCGC AGCGCTATCC GACGATTTCG AGAGGGCTGG ACGTCGTTGA GCGCACCTGC TGTTGCTGCT 7100
7101 GGTCCTACCG CCGCGGGGGC AACCGCTGCG CGGCCTGCCA CCACCCGGGT GACGATCCTG ACCGGCAGAC GGATGACCGA TTTGGTACTG CCAGCGGCGG 7200
7201 TGCCGATGGA AACTTATATT GACGACACCG TCGCGGTGCT TTCCGAGGTG TTGGAAGACA CGCCGGCTGA TGTACTCGGC GGCTTCGACT TTACCGCGCA 7300
7301 AGGCGTGTGG GCGTTCGCTC GTCCCGGATC GCCGCCGCTG AAGCTCGACC AGTCACTCGA TGACGCCGGG GTGGTCGACG GGTCACTGCT GACTCTGGTG 7400
7401 TCAGTCAGTC GCACCGAGCG CTACCGACCG TTGGTCGAGG ATGTCATCGA CGCGATCGCC GTGCTTGACG AGTCACCTGA GTTCGACCGC ACGGCATTGA 7500
7501 ATCGCTTTGT GGGGGCGGCG ATCCCGCTTT TGACCGCGCC CGTCATCGGG ATGGCGATGC GGGCGTGGTG GGAAACTGGG CGTAGCTTGT GGTGGCCGTT 7600
7601 GGCGATTGGC ATCCTGGGGA TCGCTGTGCT GGTAGGCAGC TTCGTCGCGA ACAGGTTCTA CCAGAGCGGC CACCTGGCCG AGTGCCTACT GGTCACGACG 7700
7701 TATCTGCTGA TCGCAACCGC CGCAGCGCTG GCCGTGCCGT TGCCGCGCGG GGTCAACTCG TTGGGGGCGC CACAAGTTGC CGGCGCCGCT ACGGCCGTGC 7800
7801 TGTTTTTGAC CTTGATGACG CGGGGCGGCC CTCGGAAGCG TCATGAGTTG GCGTCGTTTG CCGTGATCAC CGCTATCGCG GTCATCGCGG CCGCCGCTGC 7900
7901 CTTCGGCTAT GGATACCAGG ACTGGGTCCC CGCGGGGGGG ATCGCATTCG GGCTGTTCAT TGTGACGAAT GCGGCCAAGC TGACCGTCGC GGTCGCGCGG 8000
8001 ATCGCGCTGC CGCCGATTCC GGTACCCGGC GAAACCGTGG ACAACGAGGA GTTGCTCGAT CCCGTCGCGA CCCCGGAGGC TACCAGCGAA GAAACCCCGA 8100
8101 CCTGGCAGGC CATCATCGCG TCGGTGCCCG CGTCCGCGGT CCGGCTCACC GAGCGCAGCA AACTGGCCAA GCAACTTCTC ATCGGATACG TCACGTCGGG 8200
8201 CACCCTGATT CTGGCTGCCG GTGCCATCGC GGTCGTGGTG CGCGGGCACT TCTTTGTACA CAGCCTGGTG GTCGCGGGTT TGATCACGAC CGTCTGCGGA 8300
8301 TTTCGCTCGC GGCTTTACGC CGAGCGCTGG TGTGCGTGGG CGTTGCTGGC GGCGACGGTC GCGATTCCGA CGGGTCTGAC GGCCAAACTC ATCATCTGGT 8400
8401 ACCCGCACTA TGCCTGGCTG TTGTTGAGCG TCTACCTCAC GGTAGCCCTG GTTGCGCTCG TGGTGGTCGG GTCGATGGCT CACGTCCGGC GCGTTTCACC 8500
8501 GGTCGTAAAA CGAACTCTGG AATTGATCGA CGGCGCCATG ATCGCTGCCA TCATTCCCAT GCTGCTGTGG ATCACCGCGG TGTACGACAC GGTCCGCAAT 8600
8601 ATCCGGTTCT GAGCCGGATC GGCTGATTGG CGGTTCCTGA CAGAACATCG AGGACACGGC GCAGGTTTGC ATACCTNCGG CGCCCGACAA ATTGCTGCGA 8700
8701 TTGAGCGTGT GGCGCGTCCG GTAAAATTTG CTCGATGGGG AACACGTATA GGAGATCCGG CAATGGCTGA ACCGTTGGCC GTCGATCCCA CCGGCTTGAG 8800
8801 CGCAGCGGCC GCGAAATTGG CCGGCCTCGT TTTTCCGCAG CCTCCGGCGC CGATCGCGGT CAGCGGAACG GATTCGGTGG TAGCAGCAAT CAACAAGACC 8900
8901 ATGCCAAGCA TCGAATCGCT GGTCAGTGAC GGGCTGCCCG GCGTGAAAGC CGCCCTGACT CGAACAGCAT CCAACATGAA CGCGGCGGCG GACGTCTATG 9000
9001 CGAAGACCGA TCAGTCACTG GGAACCAGTT TGAGCCAGTA TGCATTCGGC TCGTCGGGCG AAGGCCTGGC TGGCGTCGCC TCGGTCGGTG GTCAGCCAAG 9100
9101 TCAGGCTACC CAGCTGCTGA GCACACCCGT GTCACAGGTC ACGACCCAGC TCGGCGAGAC GGCCGCTGAG CTGGCACCCC GTGTTGTTGC GACGGTGCCG 9200
9201 CAACTCGTTC AGCTGGCTCC GCACGCCGTT CAGATGTCGC AAAACGCATC CCCCATCGCT CAGACGATCA GTCAAACCGC CCAACAGGCC GCCCAGAGCG 9300
9301 CGCAGGGCGG CAGCGGCCCA ATGCCCGCAC AGCTTGCCAG CGCTGAAAAA CCGGCCACCG AGCAAGCGGA GCCGGTCCAC GAAGTGACAA ACGACGATCA 9400
9401 GGGCGACCAG GGCGACGTGC AGCCGGCCGA GGTCGTTGCC GCGGCACGTG ACGAAGCGGC CGGCGCATCA CCGGGCCAGC AGCCCGGCGG AGGCGTTCCC 9500
9501 GCGCAAGCCA TGGATACCGG AGCCGGTGCC CGCCCAGCGG CGAGTCCGCT GGCGGCCCCC GTCGATCCGT CGACTCCGGC ACCCTCAACA ACCACAACGT 9600
9601 TGTAGACCGG GCCTGCCAGC GGCTCCGTCT CGCACGCAGC GCCTGTTGCT GTCCTGGCCT CGTCAGGATG CGGCGGCCAG GGCCCGGTCG AGCAACCCGG 9700
```

Figure 1 continued, page 3 of 4

```
 9701 TGACGTATTG CCAGTACAGC CAGTCCGCGA CGGCCACACG CTGGACGGCC  GCGTCAGTCG CAGTGTGCGC TTGGTGCAGG GCAATCTCCT GTGAGTGGGC 9800
 9801 AGCGTAGGCC CGGAACGCCC GCAGATGAGC GGCCTCGCGG CCGGTAGCGG  TGCTGGTCAT GGGCTTCATC AGCTCGAACC ACAGCATGTG CCGCTCATCG 9900
 9901 CCCGGTGGAT TGACATCCAC CGGCGCCGGC GGCAACAAGT CGAGCAAACG  CTGATCGGTA GTGTCGGCCA GCTGAGCCGC CGCCGAGGGG TCGACGACCT 10000
10001 CCAGCCGCGA CCGGCCCGTC ATTTTGCCGC TCTCCGGAAT GTCATCTGGC  TCCAGCACAA TCTTGGCCAC ACCGGGATCC GAACTGGCCA ACTGCTCCGC 10100
10101 GGTACCGATC ACCGCCCGCA GCGTCATGTC GTGGAAAGCC GCCCAGGCTT  GCACGGCCAA AACCGGGTAG GTGGCACAGC GTGCAATTTC GTCAACCGGG 10200
10201 ATTGCGTGAT CCGCGCTGGC CAAGTACACC TTATTCGGCA ATTCCATCCC  GTCGGGTATG TAGGCCAGCC CATAGCTGTT GGCCACGACG ATGGAACCGT 10300
10301 CGGTGGTCAC CGCGGTGATC CAGAAGAACC CGTAGTCGCC CGCGTTGTTG  TCGGACGCGT TGAGCGCCGC CGCGATGCGT CGCGCCAACC GCAGCGCATC 10400
10401 ACCGCGGCCA CGCTGGCGGG CGCTGGCAGC TGCAGTGGCG GCGTCGCGTG  CCGCCCGAGC CGCCGACACC GGGATCATCG ACACCGGCGT ACCGTCATCT 10500
10501 GCAGACTCGC TGCGATCGGG TTTGTCGATG TGATCGGTCG ACGGAGGGCG  GGCAGGAGGT GCCGTCCGCG CCGAGGCCGC CCGCGTGCTC GGTGCCGCCG 10600
10601 CCTTGTCCGA GGTAGCCACC TGCGTCCGCC CAGTGGCAGT ATGCGGACCC  CGGNNNNNNN NNNCTCGAGT GCGTTCTTCG GAGGTTTCCA ATTCTTGGAT 10700
10701 TCCAGCACCC GCTCAGCGGT CTCGGCGACC AGACTGACAT TGGCCCCATG  CGTCGCCGTG ACCAATGAAT TGATGGCGGT ATGGCGCTCA TCAGCATCCA 10800
10801 GGCTAGAGTC ATTCTCCAGG ATATCGATCT CCCGTTGAGC GCCATCCACA  TTATTGCCGA TATCGGATTT AGCTTGCTCA ATCAACCCGG CAATATGCCT 10900
10901 GTGCCAGGTA ATCACCGTGG CGAGATAATC CTGCAGCGTC ATCAATTGAT  TGATGTTTGC ACCCAGGGCG CCGTTGGCAG CATTGGCGGC GCCGCCGGAC 11000
11001 CATAGGCCGC CTTCGAAGAC GTGGCCTTTC TGCTGGCGGC AGGTGTCCAA  TACATCGGTG ACCCTTTGCA AAACCTGGCT ATATTCCTGG GCCCGGTCAT 11100
11101 AGAAAGTGTC TTCATCGGCT TCCAGTCAGC CGCCCGGATC CAGCATCTGT  CTGGCATAGC TGCCCGTCGG CCTGGTAATA CTCATCCCCT ACTGCCCTCC 11200
11201 CCAAACCGCC AGATCGCCTC GCGGATCACC GTCCGGTTGG CCTCCGGCAT  TTCACGCCGG CTCGGCCGCT GGATCCACCC CGCGCCGGTA TTCGCAGTAA 11300
11301 CCCGTTGAAT CCGCGCGCAT GATGCACCGC TTGGGCGATC AGCCGGGTGG  TCACCTCGCT TGCGCTGGCC GCGCTGTCGC ACGGGGCGCT CGGTGGTAAC 11400
11401 GGACGTCATA ATTAACCAGC GTAACCGAAC CTAAGACCAG CTAGCTGCGG  CAATATTGGC GACCAGGACT ATGGCGCCCT CCGAACCCGG CCGATCCATG 11500
11501 TCAAAACATT GACAATGCGT ACTCACGCCG TGTCGGGCGC GCTGAATGAC  CGCATTGCGG CGCTCATTCG GTGCGTAGTC GCTACCACCG CAACAATGGG 11600
11601 CTTAGGCCAT TCCTTCGTTC ATCGCGCGGG ACATGGCCGA TAACGCAGCG  GTCAGCTGCT CGCCCGCCGC GTCGTTATAC GCGGACGCCG CGGCCTGCGC 11700
11701 ATTGTGCAGC GCCTCGTTGA CCCGCTGAGC CGCCGCCTCG GCACCCAGCT  TCTTCAGCAA ACCATCTTCG ATGCGCAGGC CGGTGAGCCA CTGGTGCCCA 11800
11801 TTGATCGTCA CTTCGACGGT CTCGGCTTCG TCGGTGGCGC GGAAGGATCC  GTTGTTCATC TGATTGAGCG TCCCGTCTAG GGCCGACTGA AACCGCGCCG 11900
11901 CCAGCGTCAA CGCCCGGGCG ACATGCGGGT CCAATTCGTC CATGCTCACT  TCGACTCCTT ACTGTCCTGG CGCCGACGGT TACCAATGAC GGCCTCGGTC 12000
12001 CATGCCCGAT CCTCGGTGTA GAGCGCCTCG TCTTCCTGCT GAGAACCCTT  GGACTTGGCG CCCCCTTGTC CCTGATGCGC GGCACCCATC GGCATTCCCA 12100
12101 TGCCACCGCC GCCCAGCGCG GCGCCGCCGC CGGCCCTTCC CTGGCCTAAG  CCGGCAATGT CACCAGCGCC AGCGGGCCGC ACCGATTCGG CGCCCCCGAT 12200
12201 CGCGGATCCC AACGGCGCCG ACGGCACCCC GCCGCCTCCA CCGCCACCGA  GCGATGCCGC TTTGACCGCC ACGTCGCCCG ACAGCGCTGC GGCTTCCCGC 12300
12301 CCAGCCGACG TCAGCTGCGC CGCCGTGTCA GCCGGGAGGC CACCACCCGG  CGATCCGGTA GGCGGAACCA TCGGTGCGGC TGGCATCCCG GTACCGGGAG 12400
12401 TCACACCGGA GCCGTCAGAC GGCGGCATCA GGAAGCCAGG GATCAATCCC  TGCTCTTGCG GAGGCGGGGC GGGTCGATCT TGATGGCGGG GGGGAGGCTT 12500
12501 CGGCGGGTTT ACCGGTTCCA GGGCTGCCTT GTTGTTGTAT TCGGTCAGCA  CCTTCTCCGA CCTCTGCTGA TACTCCGCGT ACACCGGGAG AATTTGGTCG 12600
12601 CGGGCCGAAG GGTTTTCCGC GTAAAGCCGT TCGAGCCCGA CTATGTCTTC  ATAAGTCGGA TGTTCCCGCC TAGCCCACAC GTGCAGCTGC GCGACATATT 12700
12701 GAGCCTGCTT GGCCATCGCA GCGCTCAATT TGGCCATGTG GAGTATCCAT  TGCCGGTGTT GATCGAGCGA AGCCTCGCAA GCGGTAGCCG CATCGCCTTC 12800
12801 CCAGTTGTCA AACCCCCGGA ACCGCTTGAC GTCGCCTTGC AGCGTCAGGT  TGAAAGTGTT CCACCCATCC GCAAAGTGCG CGAGCGATGC GCCTTGGTCG 12900
12901 CCCGTTTCGA GCTTCCTTGC CGCTTCTTTG AGATCCATGA AGTTGGGTTC  ACCGGCCGTG GCCACCCTCG GCGTATCGGT TAGTTCGGCC GAACTGTCCC 13000
13001 CTCCGACGGC CCCGGCCGAT TCTGCCTGCA CAGTTCCTTC GCCGTCGTTG  TCCAGCGCGG TCGCAGCCTC CTCATCAACC TCGCCATACG CCTTGGCCGC 13100
13101 GTTCGCGCAGC GAGGTCGCCA GACGCTGCCG CTCTTTGGCA CCGGCCGCCA  GGTATTCCCG CATGTTGTCG GCGGACAATA CCAGCTGTTG GGCGGCGTTT 13200
13201 TTAGCCGCCG TGAGTTCGCA CGGTGTGATG GGGACATCAG TCGGTGGGTC  CGCCATCGGG GCCTCCACCT CGTTGGCCCT GTTCAAAATC TCTTGCTGAT 13300
13301 CCACCGTCAC GGTCTGCGAC TGCGTCATAT CGGATCATCC TCCTTAGTGC  TATAGCCATT ATCGTCGCTA AACTGAAAGG TTCCTGCACT AATTTGATGC 13400
13401 CGCCCGTTCA TGCCGGCATC GCGAACGGAT CGCCCTACTT CGGCAGCGCC  ATCTGGTAGC GGCTTTCCTC GGGTGGGGAA ACCCGGCGAA TCGGCAGCTG 13500
13501 CCGATGCCGC GGGGTACCGA TCACATTGTG CCGCAGAATC ACCCGGTCAA  TACCGGGATG CGGGCCGAGA TAGGTCGTCG CATTCGGCCA CGCCACCTTT 13600
13601 ACCTCCTGCC CGATGTGTGC GCCGATCAAC CGGGCAAATT CCTCGAACTG  TGGCCCGACT GTGACCATCG CACCTGCCGC CGCCGCACGC ACCACGAACT 13700
13701 GGGTGAATGT CTGAGCGTCA CCCAGGTTGA GGGCGATGTC GACATCGTCG  AAGGGCATGT AGACCGGGCA TCGGTTCACC GTCTCGCCGA CCAGTACCCC 13800
13801 AGCTGACCCG ATCGGCAGCT GGCAGTGGCG GTTGGCCACC AGATGCTGGC  CTTGCAGCGC GGGCCGCTGC CCGCCAAATA GGCGGGCGAA GCCCCTGGGT 13900
13901 GTCTTGGGCT TGTCCGCCGT GGTCAGCAAC ACCGTGGACT GCGGGGCCAT  CCCCGGCGCG ACCCGGACTC TGGTGATGGT GTGGTCCGCG CGCGCCGACC 14000
14001 ACCATACATC CGGACCTCCG GGCGCCGCGT AGGCGGCAGT GTAGGCATCG  CGCCCCTTGA TCATCGACCA TTTCTCCCGC ACAAAGCCGA TGTCGGTGGC 14100
14101 GTGGTCGTAG TCATCGAAGC TGCGGCCACA CACCGCGTCG ACACCATGGC  TAGCCAGTCG ATCGGCAATG CGCGTCGCGG ACGCCACCAA ATACCGGGCC 14200
14201 AGTCCTGCGA CGCCTTCATC GCGGCGCTGC GCCGATTTGC GGGTGCGTTC  CGGGTCGGCG CGCAGCACGA TCCAGGTCCG GCGGTTCGCC GGCGCCGGGT 14300
14301 CTGTCCCGAT CACCTGCTGA TACAGACTCA CCACGTCCGG CGCTGCGGTA  TTGCCGACGC GGTAGCCGGC TGAGACGATA TCGGCCTCCA AGTCGGGACA 14400
14401 GTGCACCGAC AGGAGCTCCT CCACCAGTCC GGTGTCCAGC ATGTCGTCGG  TGTGGGCTTG CCCGTCGACG ATGACCGTCG GCGTGAATGG TCGGGGAATG 14500
14501 AGCTCGATTA CGGCGACCAG AAACTCGCCT TGCCAGCGCA CCGCAACGTG  ATCTCCTGGC TTCACGGTGG CCCCGACCAC AGGTTCTGAC GAGGAATCCG 14600
14601 GGGGCCGTCG GCGCCGCCGC AACCACGCGT ACACCGCCGC CACCCAGCCG  GTGATCCGGC GGCCGTAGAA AGTGACCGTG GCCACGATGA CGCCCAACGA 14700
14701 GGCCAGCGCA ATCCCCGCCC ACCAGTAGCG CGTCTCCAAG AATGCGATGA  TGCATGGCGG GGCCAACGCG GAGGCAAGCA AGGCGTGCCC GGTCCTGAAC 14800
14801 CGCAGCCCTA AAGGATTTCT CATCGGCGGC TCAGCGCCCG TCTAGCCAGC  GCGCCCAGGC CCAGGGCCAA CGTAAGGCCG ACGGCCACCA ACGCCACAGC 14900
```

Figure 1 continued, page 4 of 4

```
14901 CGTAATCGGG CGACGATCGG GACCCGGCTC CACCACCGGG GGTGGAAGTC  GTCTGACGTT GTATGGCGCC GAAGCAGGGC CGGGCGGAAT GTCCCACGTC 15000
15001 AGCGCGGCCA CCGCATCGAT GACGCCGGCG CCGACCAGGT CGTCGACCCC  GCCCCCGGGG TGTCTCGCGG TGGCGGTGAT CCGGTGGATG ATCTGCGCCG 15100
15101 GCGTCAGGTC GGGGAACCGC TGCCGAAGCA GGGCCGCCAG ACCCGACACA  TATGCCGCGG CAAACGAGGT GCCGGCGATG GGTACCGGCC CCTCCCGGCC 15200
15201 TTGCAACGCA TTCACCGGTT CACCGGTGTC GCCGAGCGCG ACGATGTTTT  CTGCGGGCGC GGCCACGTCC ACCCACGGTC CGTGCATCGA GAACGAGCTG 15300
15301 GGCATCCCGG TCTGGCCGAT ACCGCCGACG CTTAACACCA GCGGTGCGTA  CCACGCCGGG GTGACAACGG TCTGCACATT GTTCCAGCCG CGTGGGTCGC 15400
15401 CGGGTGTGGA CGGGTCCGGC GCCGGATTCT GTACGCAATC GCCACCGGTG  TTGCCGGCCG CGACCACCAC CACCACGCCT TTGACGTTGA CCGCATAGTC 15500
15501 GATGGATGCA CCCAGTGAGG TTTCATCGAT CGGCCTGCTC ACCTTGTAGC  AGGCGGCTTC ACTGATGTTG ATCACACCCA CGCCGAGGTT GGCGGCGTGC 15600
15601 ACCACGGCGC GGGCAAGACT GCGGATGGAA CCGGCGGCCG GGGTGGCGTT  GGGGTCATTC GGGTTGGCTT GTGAGCCGAC CGGTTCGAAG GCCTCAGACG 15700
15701 TCTGACGTAG CGAGAGCAGT CGAGCGTCGG GCGCGACGCC GACGAACCCG  TCGGTGGGCG CGGGCCGGCC CGCGATGATG GATGCTGTGA GAGTCCCATG 15800
15801 GGCATCACAG TCAGACAGGC CGTTACCGGC CTGGTCGACG AAATCGCCGC  CAGGTTCCGC CGGGACCCGT GGCGAAGCGT CGACACCGGT GTCGATCACC 15900
15901 GCCACCGTCA CCCCGGCCCC GGTCGCGAAC TTGTGGGCAT CGGCCACGCC  CAGATACGTG TTGCTCCACG GCGGATCGTG GAACCCGGAC CCCGGCAGCG 16000
16001 TGGTGGGCGA CGCGCACAAA ACGCGCTGTT CGGTAGGCTG ATCCGGGCCC  GCCACGTCGG GCGGCAACGC GCCCGGATCG ATCGGCGGTG GCGTGATGGC 16100
16101 CGATGCGGGC GACGCGGTGA GCAACGCCAG CGCCACCGTG ATCAGAAAGA  TACGGTGCAC TCCCAGAACA CTCCATTCGT TGAGATTCAT TGCGATTCAT 16200
16201 TGAGCTGCGT TGCTACCTTG GGCCACTTGA CGGACCTGTG TGCATTTTAG  ACGTAACGGC TGGGCAAACA ACGCTGTCAC GCCTGGGCTG GTCCGCCGCG 16300
16301 CCGACCAGGG CGCGTAGGCG CTGTACCTGG ACCACGCCGG GACTCAACGG  TTTTGCTACC GCACTAGCCG ATATGCGGCT GCTACCAAAC GATCGCGGCC 16400
16401 ATGTCTCGGT TGTCTGAGCA CACGCTGCGT ATCGCGGCAT CGATGTCGGT  GGCGGTGATG ATCTGCAGAT CCTGAACCGA TACCGGTTGG CCCGCACGTT 16500
16501 TTTGCGCAAC CACCCGGGTG TCCCGGAACC CTTCGGCGCG TTCGATCACG  TTGCGGGCGA ACCGACCGTT TTGCATAGCG TCGATACCGT GCTGCCCACT 16600
16601 AGGGGTGGTG TAGTTACGGA TGGTGGTGAC CGCGTCGAGG AATACCTCCC  GTGCGGCGTC ATCGAGCTGG CTGGCGCGCG GTGTAGCGTA GCGGTGTCCA 16700
16701 ATCTCGACGA TCTCCACCGG CGAATAAGAC TCGAACCGCA GCTTTCGGTT  GAACCGGCCA GCCAAACCCG GGTTCACGGT GAGGAATTCG GTACCCCGGG 16800
16801 TTCGAAATCG ATAACTTGGA TCCGGAGAGC TCCCAACGCG TTGGATGCAT  AGCTTGAGTA TTCTATAGTG TCACCTAAAT ACTTG                 16885
          |  10      |  20      |  30      |  40      |  50      |  60      |  70      |  80      |  90      |  100
```

Figure 2


```
      |  10      |  20      |  30      |  40      |  50      |  60      |  70      |  80      |  90      | 100
    1 GGATCCTCGG ACTGGCCGCG GTCGTGCTTG TGCACGAGTT CACCGAGGTC ATCGTCATCG CCAACGGCGT GCGGGCCGGA CGCATCAAAC CACTTGCCGG 100
  101 GCCACCCAAG ACACCTGATC GGACTATCCC GGGGTAGCGA CGCGCGGAAT CGTGGAGTGT GTTTGGACCA GCAATAGCGT CACTGTGACG AAACAGCCGC 200
  201 CGTCTTCTGG AAGTTATACC CGGTTATACT ATCTGTATGA AGACAGCTAT TTCTCTGCCG GATGAGACGT TCGATCGGGT ATCGCGGCGT GCGAGTGAGC 300
  301 TCGGCATGAG TCGGTCCGAG TTCTTCACGA AGGCTGCGCA GCGCTACCTG CACGAGCTGG ACGCCCAATT GCTCACGGGC CAGATCGACA GGGCTCTAGA 400
  401 GAGCATCCAT GGCACCGACG AAGCGGAGGC CCTCGCCGTG GCCAACGCAT ACCGCGTGCT AGAAACCATG GACGATGAGT GGTGATTAGT CGTGCCGAGA 500
  501 TCTACTGGGC TGACCTCGGG CCGCCATCAG GCAGTCAGCC GGCGAAGCGC CGCCCGGTGC TCGTAATCCA GTCAGATCCG TACAACGCAA GTCGCCTTGC 600
  601 CACTGTGATC GCAGCGGTGA TCACGTCCAA TACGGCGCTG GCGGCAATGC CCGGCAACGT GTTCTTGCCC GCGACCACAA CGCGACTGCC ACGTGACTCG 700
  701 GTCGTCAACG TCACGGCGAT TGTCACGCTC AACAAGACTG ACCTCACCGA CCGAGTTGGG GAGGTGCCAG CGAGCTTGAT GCACGAGGTT GACCGAGGAC 800
  801 TTCGTCGCGT ACTGGACCTT TGACACTGCG CCACGCGACA ATTCGTCACG GTGACGTTCC TGCTTGGTGT AAGCCCCCCC GCCGGGGGAA CTACTCGCCG 900
  901 GAGGTGGTGT TGTGGGCAGG CTTGAGGGCA AGGTTGCATT CATTACGGGC GTGGCTCGGG GTCAAGGCCG TTCGCATGCG GTCCGCCTAG CCGACGGCCA 1000
 1001 AGCGCGTGCG CTCGGCAAGG TCGATGTTGA GGCGTGCGGT GCGCTCGTTG GTGAGGTAGA AGTGTGGGGC CGTGACGTGC GTGACGATCG ACGGGTGTTT 1100
 1101 GTCGAGAGTC CTGCCGACGA GTTCGGCGCG TGCCGCCGCG TCGCGCGTCA GGGCATCCGT GTCGTAGGGC TGCCCGTTTC ACAGAGGGAA CTTGTCGAGC 1200
 1201 CCGAAGCCGG GTGCGCGGCG AGGCGCTCGG CTGCTGGCTC CCAGTAGACA TCTAGGCCTG CGTCGACTGC GGCTGCGGCA GCGTCGTGCT GGTGACGAGT 1300
 1301 GGCGTTGGTG TCCAGCGTGA TCGCAGTGGT GCCGGCGTGG TCGCGGGACA GGAAGTCCTC GACCGGTTTG TGATCACCCG GCCCGAGCCG AAACTGAATG 1400
 1401 CCCATCGTCG TGAAGTTCCT CTCGCATCGA CGCCTCGGTT CGTGTCATAA TACATGACAA ATCAATAGAC AAAAGGAAGA CAGGCTGCCC ATGGGAGTAA 1500
 1501 ATGTGCTCGC CTCGACCGTG TCGGGTGCGA TCGAGCGCTT GGGATTGACC TACGAGGAAG TCGGTGACAT CGTCGATGCC TCGCCGCGTT CCGTGGCGCG 1600
 1601 ATGGACCGCA GGTCAGGTGG TTCCCCAACG CCTCAACAAG CAACGACTTA TCGAGCTGGC CTATGTCGCC GACGCCCTCG CGGAAGTGCT GCCGCGTGAC 1700
 1701 CAGGCGAACG TGTGGATGTT TTCGCCGAAT CGGTTACTGG AACACCGCAA GCCTGCCGAC CTCGTGCGAG ACGGCGAGTA CCAACGCGTG TTGGCGCTCA 1800
 1801 TCGACGCGAT GGCGGAGGGA GTGTTCGTGT GAGCGATGCC CTCGATGAAG GGCTCGTCCA GCGTATCGAC GCACGCGTAA CAATTGAGTG GTCGGAAACG 1900
 1901 TGCTACCGGT ATACCGGCGC GCACCGTGAC GCCTTGTCCG GTGAGGGCGC GCGCAGATTC GGAGGCAGGT GGAATCCGCC GCTGCTCTTT CCGGCGATCT 2000
 2001 ATCTTGCTGA TTCCGCCCAA GCCTGCATGG TTGAGGTGGA ACGGGCGGCG CAAGCGGCTT CAACGACCGC AGAGAAGATG CTCGAGGCGG CCTACCGACT 2100
 2101 ACACACGATC GACGTCACGG ACCTGGCCGT CCTCGATCTG ACAACCCCGC AAGCTCGGGA AGCCGTGGGG CTCGAGAACG ACGACATCTA TGGCGACGAC 2200
 2201 TGGTCAGGGT GCCAGGCGGT CGGACATGCG GCCTGGTTCT TGCACATGCA AGGTGTCCTC GTGCCGGCGG CGGGCGGTGT CGGCCTCGTT GTCACCGCGT 2300
 2301 ATGAACAGCG AACTCGGCCG GGCCAACTAC AACTGCGACA AAGCGTCGAT CTGACGCCTG CTCTTTACCA AGAACTTCGA GCCACGTAGC TGGCCAGCTT 2400
 2401 GGCGCAGAGA AGGATGCCGC TGTGCCATGG TCATCGTAAG GAGCAACTCG CATCACTTAT AAGCCGATAA GCGACATTAT GTCAAGTGAA GCTGGTCGTA 2500
 2501 TTGGGTTAGC TGCGCCGTTT GTGCTAGCGG GCACGCTCCT TGTGCGTGCT GCGGCAGCGA GCGTGTCGTC AAAGGTTGCG AGGCTTGCCT GGTGATGAAT 2600
 2601 TGCCACATCC GGCACGCAGC AATCAGGTAG TTTTAGCCCG CTGCTAGCGC GTAGTTCGGC GAGACGCAGC AGCTCGCCAT CGTCGTACGG CGCAACGACG 2700
 2701 CTACCCGCGG AACGCGGATC GTCGAGCATC GATGATCAGG CGCCGCAATA ATTTGGAACG GGGCTCGCCA GGCCATCGCT GGGCGGCCCG GTCCAATGCC 2800
 2801 TGAGCTACCT CCGGGGTTTC GGTTATTTGG TAGCGCGGAC GAGTGGTCGA CATAGAACGA AGTGTGCCAC TTCTAGCAAA GGTGGTACAC CTACTGGCGG 2900
 2901 CCGCGGGTTT ACCGCCCCTG CCAGTCACCG CACTTCCGGC GGCGGACGAC GAGCACCGCG GAATCCACAT GCGGCGGTGG CAGGAACGCG CGCCGTGGCA 3000
 3001 GCATGAGGCC GACGGTCAGG GTGAACCTTC GCGCGTTGCG AGAAGCGAAT TTACATACGA GGGCTCGCTG CAGCACGAGA TCGGCCGCGA CAAGCCCGCT 3100
 3101 GTTGGTGCCC AGCAGCGTCC GCAGCAGGCG GGACGAAATC CCGTACGGCG GGTTCGCCAC AACCCGGAAC GGCCGGCCGG GCAACCGGAT CGAGGCGGCG 3200
 3201 TCCGCGTGCA CCACGGTAAT GCCAGGGAAT CGCTCGCGGA GGACACCGAC TCGTCGCGGG TGCAACTCCA CGGCGACCAC CCGCGCCCCC GCTCGCACTA 3300
 3301 GATGCGCCGT CAGTGCCCCT TCGCCGCGC CGATGTCAAA CACGAGCTCA CCGGACCGCA CTGCGGCCGC GCTGACTACC CGCGCTGCCC ATTCGTCATG 3400
 3401 GAGCCGTGC CAGCCCCATG CCCGTCGCGA CCGTCCGAGG GCGGACACGA CGTACCGTCA CTGCGTAGAT GCCCACGCGC CCGACCGTAG CCCGCCACCG 3500
 3501 GCACTGCGAT CAATCCAATT TCTCGGTTCA GGCAACCTTC TGGTCATCAC CAGCCCCAGG GCTCTGGCGC CGTCCGCATC AACTCCGAGA TGACGTTGGC 3600
 3601 CGTGACGACC CACTAGACCC ACCTGGCAGT AGCCGCATTG TCGCAGTCGG CGAGCCTCAG TGCGCAGTCG CGTCTAGGTG CAAGGATATT GCCCGTTGAG 3700
 3701 CAGACAACTC GACGGCGGCG AGTAAGAACC GGTCAGCCCG CCTCTTAGGC CGCCCGTGGC TGAACCACCG GGGGCAATGA TGCGATTCCA ATTCGCTGGG 3800
 3801 CTGAGAACGT AGTGCGTGCC AGATCGTGCA ACGGTGCTAT TCCATGTGTG CAAGACGGAT TCTCCTGCCG GCAAGTCGAA TTCAAGCTTC CAATCGGTTA 3900
 3901 GCGGCGCCGT GCTCGAGTTT GTGATGGTGA AGCGGGCGAT GAAACCGGTC TGCCACGTCG ATGTCACCGA CAACGTCGCC CTGGCCGTCG CCGCACTAGC 4000
 4001 GACCGGGGTG ATGGCGAGTC CGAGGATGGC AACTATCAAT GCCGACACGG TTGCGTGAAG CGCTGTCCGC CAGCGCCTCA CGTAAATGTT CAGTCCGGCC 4100
 4101 ATGACAGCCA ACACTAATGC CAATGAGGCG ATATCGGCCG TCTCCTCGCG AGCAAGCTAC AGCAACTTTG CTCAACCGCA ACCGTGATGA AATTTGGCCT 4200
 4201 CGACCCACCC TGAACCAGAT ATCGGCCCGG CCGAACGCGA ACTTGCGGAC GGGGAAGGCC AGACAGCCTC GACCCCACTC CCCCGATTAG CGCCGTTCAC 4300
 4301 CGTTCCGCAC CGGTATCAAC GGGCTACAGC TCCAACACGA TCCGTAGGGC CGCGTCACGC CGAATGTGCA CTGGTGGCGC CGACACGCCC GGGCGAGGCC 4400
 4401 GCCGTCGGCG TGTCAGCTGG TGACTGAGTT GTGCAGACTG ACCGCGCGCC CTCCTGCCGA ACGGTATGTG CCCATCGACG ATCACGTGGT CCAACCCGCG 4500
```

Figure 2 continued, page 2 of 4

```
4501 TGTGCACACG TGCTGTACTA GGTCACGGTC AGCGAGATTC CCAGCGCAAC CATCATGACC GCGATCAGGC CGTCGAGGAT TCTCCACGAG CCGGGGTTGG 4600
4601 TGAACAGCCC GCGCAACCGG CCGGCTCCGA ACCCGAGGGT GGCGAACCAT ACCGCACTGG CTGTGACCGC GCCGAGGCCG AACAGCCAGC GCTGGTCGCT 4700
4701 GTGCTCGTTG GCCAGCGCGC CTAGCAACAC GACGGTGTCG AGGTAGACGT GTGGGTTGAG GAACGTGAAT GCCGCACAGG TCACCAGGAC CTCGGCTAAG 4800
4801 CGAACCGGCG TGGCGCCAGA TGGGATCAGC GCAACAGGTC GCCACGCCCG CCGGGCCGCA AGTAGCCCGT AGCCGATTAG GAAGGCGGCG CCGCCAAACT 4900
4901 TGACGACATT GAGCGCACGC GGATGTGCGC CGATCAATGC GCCGAACCCC GCGATACCGG CGGCGATCAG CACGATGTCG GACACCGTGC ACAGCGCCAC 5000
5001 CACCGGCAGC ACGTGCTCAC GCTGGATTCC CTGCCGCAGC ACGAATGCGT TCTGCGCGCC AATCGCGGCG ATCAGCGTGA AGCAGGCCAG GAAGCCGACG 5100
5101 ACCAGTGGTG AGTTCACGCA ATCGACACTA GGCAGTTTGT ATGGGTCAGT ATAGCTAATA ATTCTTCATT TACATTAGCA TTATTAATGT GCAGTGCGAC 5200
5201 GCTCCGCAGA TGGTCTACAC CTGAGATGGT GGATCCGCAG CTTGACGGTC CACAGCTGGC CGCATTGGCT GCCGTGGTCG AACTGGGCAG CTTCGATGCG 5300
5301 GCCGCGGAGC GCCTACATGT CACCCCCTCG GCTGTCAGTC AGCGCATCAA GTCGTTGGAG CAGCAGGTCG GCCAGGTGCT GGTGGTCAGG GAAAAGCCAT 5400
5401 GTCGGGCGAC GACCGCAGGT ATCCCGCTGT TGCGGTTGGC CGCGCAAACA GCGTTGCTCG AGTCCGAGGC GCTCGCTGAA ATGGGTGGCA ACGCGTCGCT 5500
5501 GAAACGCACG CGGATCACCA TTGCGGTAAA CGCCGATTCC ATGGCGACAT GGTTTTCGGC CGTGTTCGAC GGTCTCGGCG ACGTCCTGCT CGACGTTCGG 5600
5601 ATCGAGGACC AGGACCATTC CGCGCGGCTG CTACGGGAGG GTGTGGCGAT GGGCGCGGTG ACCACCGAGC GGAACCCGGT GCCGGGCTGC CGGGTGCACC 5700
5701 CGCTGGGTGA AATGCGCTAC CTACCAGTGG CCAGCAGGCC ATTCGTCCAG CGCCATCTAT CCGACGGGTT CACTGCCGCC GCGGCGGCTA AAGCTCCGTC 5800
5801 ACTGGCGTGG AATCGTGACG ATGGGCTGCA GGACATGTTG GTGCGTAAGG CCTTTCGTCG CGCCATCACC AGACCGACGC ACTTTGTCCC GACCACAGAG 5900
5901 GGCTTCACCG CCGCAGCGCG CGCCGGGCTG GGATGGGGCA TGTTCCCCGA GAAGCTGGCA GCATCTCCGC TTGCCGATGG ATCGTTCGTA CGGGTCTGCG 6000
6001 ACATACACCT CGACGTCCCT CTCTATTGGC AATGCTGGAA ACTGGACAGT CCGATCATCG CGCGAATTAC CGACACGGTG AGGGCGGCGG CAAGCGGTCT 6100
6101 GTACCGGGGC CAGCAACGCC GCCGCCGACC GGGTTGACCG ACGCCAGCAT GTTGTTGTGT CAGCGCGGCT TGGTCTCGAT GTCCCGGCCT TGCTGGACCC 6200
6201 GCTTCCTCAA ACAGGTTGAA CTTAACGACT CAGACGGAAA CGCTTGAACC GCGACGTCGC TCCGGACACC AATTTGACTC GGCTCTTTGG CAATTGAAGG 6300
6301 TGAGCTGCGA GCAGCCGGGT GACCGCATCG TTGGCCTTGC CATCAATCGC CGGCTCGCGG ACGTAGATAA TCAGCTCACC GTTGGGACCG ACCTCGACCA 6400
6401 GGGGTCCTTT GTGACTGCCG GGCTTGACGC GGACGACCAC AGAGTCGGTC ATCGCCTAAG GCTACCGTTC TGACCTGGGG CTGCGTGGGC GCCGACGACG 6500
6501 TGAGGCACGT CATGTCTCAG CGGCCCACCG CCACCTCGGT CGCCGGCAGT ATGTCAGCAT GTGCAGATGA CTCCACGCAG CCTTGTTCGC ATCGTTGGTG 6600
6601 TCGTGGTTGC GACGACCTTG GCGCTGGTGA GCGCACCCGC CGGCGGTCGT GCCGCGCATG CGGATCCGTG TTCGGACATC GCGGTCGTTT TCGCTCGCGG 6700
6701 CACGCATCAG GCTTCTGGTC TTGGCGACGT CGGTGAGGCG TTCGTCGACT CGCTTACCTC GCAAGTTGGC GGGCGGTCGA TTGGGGTCTA CGCGGTGAAC 6800
6801 TACCCAGCAA GCGACGACTA CCGCGCGAGC GCGTCAAACG GTTCCGATGA TGCGAGCGCC CACATCCAGC GCACCGTCGC CAGCTGCCCG AACACCAGGA 6900
6901 TTGTGCTTGG TGGCTATTCG CAGGGTGCGA CGGTCATCGA TTTGTCCACC TCGGCGATGC CGCCCGCGGT GGCAGATCAT GTCGCCGCTG TCGCCCTTTT 7000
7001 CGGCGAGCCA TCCAGTGGTT TCTCCAGCAT GTTGTGGGGC GGCGGGTCGT TGCCGACAAT CGGTCCGCTG TATAGCTCTA AGACCATAAA CTTGTGTGCT 7100
7101 CCCGACGATC CAATATGCAC CGGAGGCGGC AATATTATGG CGCATGTTTC GTATGTTCAG TCGGGGATGA CAAGCCAGGC GGCGACATTC GCGGCGAACA 7200
7201 GGCTCGATCA CGCCGGATGA TCAAAGACTG TTGTCCCTAT ACCGCTGGGG CTGTAGTCGA TGTACACCGG CTGGAATCTG AAGGGCAAGA ACCCGGTATT 7300
7301 CATCAGGCCG GATGAAATGA CGGTCGGGCG GTAATCGTTT GTGTTGAACG CGTAGAGCCG ATCACCGCCG GGGCTGGTGT AGACCTCAAT GTTTGTGTTC 7400
7401 GCCGGCAGGG TTCCGGATCC GATGACATAT GACGGGATGG TTCCCGTTAC CCCACCGGAA TCGATGATCG AGGGGACGGG TATGGGAGTC CCACCATCGA 7500
7501 TCTTTACGTA CAGGGTGGTG ATCGGCGATC CGACGACCTC GACGTTGGGC GCAGGTAGCG GGTTGGGACC GAACACGAGC TCACCTGCGG GTGCGTCGAT 7600
7601 GAGCACTCCC TGGTTGAGGT CACCCGGTAA CGCCATCGTC GGAATGCTGG GGCCTGGTCC CACCGCATTG GACCCAACTC CCAGAACGCC GTCCACGCCG 7700
7701 ACGGCACCGA AATAGGCTTC GAACGGGGTT GTTGTCGGAT CGGCCAGCAA GGCGCTGAAG TAGGTCGAAA TGGCGAAGGG GGACGTTGGG ATGGACAAGA 7800
7801 GGACGACATT AACGGTGGTC GGCGCGGTGA CGATGCCATT CCCGAAGTCC ACCGTCGTGG TATACGTGGC GAAGATGTAG TACAGCCCCC CGCTGTAACC 7900
7901 GCTGATGCTC AATCCGGTTG GGAGGCCCAT GTGAAGCACT CCCAGGATTC CCCCGACATC CTCAGGTGAG ACAACAAGAT CAGCGGATCC GGTGTCGACC 8000
8001 AGAATGGTTG ACGTCGGTCC GCCGTTGACG TTGGCATGTA CCGTCGGCTC TGTGACATGA ATTATCTCCA GCGGGACGGT TCTGCCGTCG CCGCCTGGCC 8100
8101 CACTGACGCC GTATCCGCCA TACAACAGGC CGCCGCGGCC TCCGGCGCCG CCCGTGCCAC GATCGACACC GACTCCATCA CCGAGGCCGC CGGCTCCGCC 8200
8201 GTTCGCTCCC CAGCCCAAAA GTCCTGTGGC ACCGCCGGTT CCGCCGGCTC CACCACTCAT GCCGGGGATG CTCGACGCGC CCCCGGGCCC GCCGATCCCG 8300
8301 CCGTTGCCGA GCAGCCAGCC ACCGTTTGCC GCCGGCTCCC CCGGGGGCGT TGGGGCCGCC GGCTCCCCCG GCGCCGCCAT TGCCGATCAA CGCCGCGGAA 8400
8401 CCGCCGGCAC CGCCGCCGAC CCCCGACGCC GTCGCGGAAT AACCGTTGCC GCCGTTGCCG TACAGCAGCC CACCCGCCCC GCCATTCGGA CTCGTTGCCG 8500
8501 TCCCGGGTGC TCCGTCGCCG ATTAGCGGAC GCCCCAACAA CGTTTCGGTT GGTGCATTGA CCGCGCCCAG CAGATCGTGC TGCGCGGTCT GCAACTGTGA 8600
8601 CGCGATGGTG GCCTCCGCGG CCGCATACGA TCCTGACGCC GAGTTCAGCG TCTGCACGAA CTGTTGATGG AAAGCGCTCG CCTGCGCGCT GACCGCTTGA 8700
8701 TATTCCTGAC CGAACCTGGC AAACAGCGCT GCCACCGCCG CCGATACCTC ATCAGCGCCA GCGGCCGCAA GCGCGGTGGT CGAGGCGGCA GCCGCGGCAT 8800
8801 TCGCCGCGCG CAGTGTGGAA CCTATGTTCT CCACATCCGC TGCCGCGGAC GTCAAGAACT CGGGAACCAC GACCAGAAAT GACACGCCGC CCCTCCGCCT 8900
8901 CGATCACCAT CCCTGCGCGC ATACAGCGTA TCCAGACGCT GCCTTTGACA TCTCGGATTT TCAGTAGCTA CCGCACGGCA CAGCACGCGT TAGGTAGATA 9000
9001 GTGGCTATTT GCTGGTACCA TCTACCTGTG GCGCTGAATA TCAAAGACCC TGAGGTAGAC CGACTAGCCG CCGAACTCGC TGACCGGCTG CACACCAGCA 9100
9101 AGACTGCCGC CATCCGGCAT GCCCTGTCTG CCCAGCTGGC GTTTTTGGAG TCGCGCGCCG GCGACCGTGA GGCACAACTT CTCGACATCT TGCGTACCGA 9200
9201 AATCTGGCCC CTGCTTGCCG ACCGCTCCCC CATCACCAAG CTCGAGCGCG AACAAATCCT CGGCTACGAC CCCGCAACCG GAGTCTGAGC ACCGCAATGA 9300
9301 TCGTGGACAC AAGCGCCGTG GTGGCCCTGG TTCAAGGCGA GCGGCCGCAC GCCACCCTGG TCGCGGCCGC CCTGGCCGGC GCCCATAGCC CCGTCATGTC 9400
9401 TGCACCCACC GTCGCCGAAT GCCTGATTGT CTTGACCGCC CGTCACGGCC CCGTTGCGCG CACGATCTTC GAACGACTTC GCAGCGAAAT CGGCTTGAGC 9500
9501 GTGTCATCTT TCACCGCCGA GCATGCCGCT GCCACGCAAC GAGCCTTTCT GCGATACGGC AAGGGGCGCC ACCGCGCGGC TCTCAACTTC GGAGACTGTA 9600
9601 TGACGTACGC GACCGCCCAG CTGGGCCACC AACCACTGCT GGCCGTCGGC AACGACTTCC CGCAAACCGA CCTTGAGTTC CGCGGCGTCG TCGGCTACTG 9700
```

Figure 2 continued, page 3 of 4

```
 9701 GCCAGGCGTC GCGTAACCGT ATGCGCGGTG ATCGCTGTTT GTAATGAGTT CAGCGACACG AAGAATAAAA TATCGGTAGC CGAAATCACT AAGCTACAGT 9800
 9801 GCTGGTGCAC GCCATGAAAG ACCGTCAATG ACAAGGAGGA CGGCCGAAAT GCCCAAGGAC CGACTGCCGG ACTTGACGCC CACAGGAGCG TACGCACCGG 9900
 9901 CCAACAGCGG CATGACCATG GCAAGGCAGG ACGGCCCTCG ATGACCGGCA AGCGCGTTGA GCGCGTGCAC GCAATCAATT GGAACCGGTT GCTCGATGCT 10000
10001 AAAGATTTGC AGGTCTGGGA ACGTTTGACC GGTAACTTTT GGTTGCCGGA AAAGATTCCG CTCTCCAACG ACCTGGCATC TTGGCAAACG TTGAGTTCCA 10100
10101 CCGAGCAGCA GACGACGATC CGGGTGTTCA CCGGCTTGAC CCTGCTCGAC ACCGCGCAGG CGACGGTGGG AGCAGTGGCC ATGATCGACG ACGCGGTCAC 10200
10201 CCCCCACGAA GAGCCGGTCC TGACCAACAT GGCGTTCATG GAGTCAGTGC ACGCCAAGAG CTACAGCTCG ATCTTCTCGA CCCTGTGCTC GACCAAGCAG 10300
10301 ATCGACGATG CCTTCGACTG GTCGGAACAG AACCCTTACC TGCAGCGAAA AGCGCAGATC ATCGTCGACT ACTACCGCGG TGACGACGCG CTCAAGCGCA 10400
10401 AAGCATCGTC GGTAATGCTG GAGTCCTTCC TGTTCTACTC CGGCTTCTAC CTGCCCATGT ACTGGTCGTC GCGGGGTAAG CTCACCAACA CCGCCGATCT 10500
10501 GATCCGGCTG ATCATCCGAG ATGAAGCCGT CCACGGCTAC TACATCGGCT ACAAATGTCA ACGAGGTTTG GCCGACCTGA CCGACGCCGA GCGGGCCGAC 10600
10601 CACCGCGAAT ACACCTGCGA GCTGCTGCAC ACGCTCTACG CGAACGAGAT CGACTATGCG CACGACTTGT ACGACGAGTT GGGCTGGACC GACGACGTTT 10700
10701 TGCCCTACAT GCGTTACAAC GCCAACAAGG CGCTAGCCAA CCTGGGATAC CAGCCTGCAT TCGATCGTGA CACCTGCCAG GTGAACCCGG CCGTGCGCGC 10800
10801 AGCTCTCGAC CCCGGTGCAG GGGAGAACCA CGACTTTTTC TCCGGCTCCG GAAGCTCATA CGTAATGGGC ACCCACCAAC CCACCACCGA CACCGACTGG 10900
10901 GACTTCTAAC CGCCCAGCGC GTCGGGGGCG TCGAGCACCA CGCGACACCG GGCCCGATCG ATCTGCTAGC TTGAGTCTGG TCAGGCATCG TCGTCAGCAG 11000
11001 CGCGATGCCC TATGTTTGTC GTCGACTCAG ATATCGCGGC AATCCAATCT CCCGCCTGCG GCCGGCGGTG CTGCAAACTA CTCCCGGAGG AATTTCGACG 11100
11101 TGCGCATCAA GATCTTCATG CTGGTCACGG CTGTCGTTTT GCTCTGTTGT TCGGGTGTGG CCACGGCCGC GCCCAAGACC TACTGCGAGG AGTTGAAAGG 11200
11201 CACCGATACC GGCCAGGCGT GCCAGATTCA AATGTCCGAC CCGGCCTACA ACATCAACAT CAGCCTGCCC AGTTACTACC CCGACCAGAA GTCGCTGGAA 11300
11301 AATTACATCG CCCAGACGCG CGACAAGTTC CTCAGCGCGG CCACATCGTC CACTCCACGC GAAGCCCCCT ACGAATTGAA TATCACCTCG GCCACATACC 11400
11401 AGTCCGCGAT ACCGCCGCGT GGTACGCAGG CCGTGGTGCT CAAGGTCTAC CAGAACGCCG GCGGCACGCA CCCAACGACC ACGTACAAGG CCTTCGATTG 11500
11501 GGACCAGGCC TATCGCAAGC CAATCACCTA TGACACGCTG TGGCAGGCTG ACACCGATCC GCTGCCAGTC GTCTTCCCCA TTGTGCAAGG TGAACTGAGC 11600
11601 AAGCAGACCG GACAACAGGT ATCGATAGCG CCGAATGCCG GCTTGGACCC GGTGAATTAT CAGAACTTCG CAGTCACGAA CGACGGGGTG ATTTTCTTCT 11700
11701 TCAACCCGGG GGAGTTGCTG CCCGAAGCAG CCGGCCCAAC CCAGGTATTG GTCCCACGTT CCGCGATCGA CTCGATGCTG GCCTAGACTC GCGAGGACCG 11800
11801 CGCGGTGGTC ACTGCGCGGA TTTGGGGCGG CGGAAGTGAG TGTTCGGTGC GCCCACTGCG GTGACTCACC TGCAGCGCCG GCATCGACAG GCCGGGAGCT 11900
11901 CAAGAATCGT CGCTAGAGAA TCTATGGTGC GTTAGAGGAT TCCCTGCTAG ACAGCCTTGG TGCGGTGGTC GGCCCGCGGA CGAGAGGATA TGCGATCCAC 12000
12001 AAGCTGGGTT TCTGCAGCGT CGTCATGCTC GGGATCAACT CGATAATCGG CGCCGGTATC TTCCTAACTC CAGGTGAGGT GATCGGGCTC GCAGGACCCT 12100
12101 TCGCGCCGAT GGCCTATGTT TTAGCTGGCA TTTTCGCGGG TGTCGTGGCG ATCGTCTTCG CGACGGCGGC AAGGTACGTC AGAACAAACG GTGCCTCCTA 12200
12201 CGCCTACACA ACGGCCGCAT TTGGGCGCCG GATCGGCATC TATGTCGGTG TCACCCACGC CATTACCGCG TCCATCGCTT GGGGGGTGTT GGCTTCTTTT 12300
12301 TTCGTCTCGA CGCTGTTGCG AGTGGCCTTC CCCGACAAGG CCTGGGCCGA CGCCGAGCAA CTGTTCAGTG TGAAGACGCT GACGTTTCTC GGCTTTATCG 12400
12401 GCGTGCTGTT GGCCATCAAC CTCTTCGGCA ACCGGGCGAT CAAGTGGGCC AACGGAACGT CAACGGTAGG CAAGGCATTC GCGCTCTCGG CATTCATTGT 12500
12501 CGGCGGGCTG TGGATCATCA CCACCCAGCA CGTGAACAAC TACGCAACGG CGTGGTCGGC ATACAGCGCG ACCCCGTACT CGTTGCTTGG CGTCGCCGAA 12600
12601 ATTGGCAAGG GCACGTTCTC GAGTATGGCG CTGGCCACGA TTGTCGCGTT GTACGCATTC ACCGGTTTCG AATCGATCGC GAACGCCGCC GAAGAAATGG 12700
12701 ACGCGCCGGA CCGGAACCTG CCGAGAGCTA TACCGATCGC GATCTTCTCG GTTGGCGCGA TCTACTTGCT CACCCTAACG GTAGCGATGC TGCTCGGATC 12800
12801 GAACAAGATC GCCGCGTCGG GCGACACCGT GAAACTGGCC GCGGCCATCG GAAACGCTAC CTTCCGAACG ATCATCGTCG TCGGAGCCCT GATATCGATG 12900
12901 TTCGGCATCA ATGTCGCGGC CTCGTTCGGT GCACCGCGGC TTTGGACCGC GTTAGCGGAC AGCGGGGTTC TGCCGACACG CTTGTCACGC AAGAACCAAT 13000
13001 ACGACGTGCC GATGGTCTCC TTCGCAATTA CGGCGTCGTT GGCGCTCGCA TTCCCGTTGG CGCTGCGGTT CGACAACCTG CACCTGACCG GCCTGGCGGT 13100
13101 GATCGCCCGA TTCGTCCAGT TCATCATCGT GCCGATCGCT CTCATCGCAT TGGCGAGGTC TCAGGCAGTA GAACATGCTG CTGTGCGGCG AAATGCGTTC 13200
13201 ACCGACAAGG TGTTACCGCT TGTTGCGATC GTGGTCTCGG TTGGGCTGGC AGTGTCCTAC GACTACCGCT GCATCTTTCT AGTGCGGGGT GGTCCGAACT 13300
13301 ACTTCTCGAT TGCTTTGATC GTGATCACGT TCATCGTGGT ACCGGCGATG GCTTATCTGC ACTACTACCG AATCATTCGC CGGGTTGGCG ATCGGCCGAG 13400
13401 CACTCGCTAG ATTCCGTTGG CGCTGAGCTC GAACGGGAGA ACACAACGGC GAGCGATGGC GGGAATAGCC TGGTCGGTGC GGGCAAGATT TCAACCTGCA 13500
13501 TTCCCGGATC GGCGGCGCGG GCAAGCGTCT GCAACGCCGA GGGACTGTAG GCACGTAGTG CGCTGATAAA GCCGTCGTGC ATGCTCGAGC GCATCGACGA 13600
13601 CCATGGCAGC AGCAGTAGGT GGAGCGGCAG TAGCAGCACC GAAGAGAGCG TGAACGACAG CGGTTTCTGC CGTTTGAGGT CGATGATCAG AAAGCGCTTC 13700
13701 CCCACCCGGG TGGCCTCGGC GATCGCTTTG CAGCGACCG TAGGCGGCAG GTGGTGAAAT GCCAGCCGCA AGACCGCCAG GTCATAGCTG TGGTCGTGGC 13800
13801 CGTCGATTGC GGTGGCGTCG ATCACTTGGG TGCGTGCTCG CGGATGTGTT CCCAGCTCTC CCGCGGCGAT GTTGGCCACC GAGGTGGGAT CTAGATCGCT 13900
13901 GATCGTCACC GTCGCTGTCG GGTGTAGCTC GAGGATTTTC GCTGAGAGCT TGCCATGGCC CGCACCAAGT TCCAGGATTC GCGGGTTGGG AATGTCAGAA 14000
14001 ACAAGTTTCA GGGCTATCCG GGCGTACTTC TCGTGCAGGT TGGTCAGGGT GCCCACCCGG TCGAGCACCC CGATGATCTT CTGTTTGACC TCATCGGGCA 14100
14101 CATCGTCGCG GTCGAGGTAC TCCAGTGCGT CGGTCTGGAA TCGACGATCC AGCCAAGACG CGTCGGGGCC ACCCCGTGGC ATCGTGGCGA TCGCCTGCTC 14200
14201 GCGGATGTTC GCCTCACCCA TGGCAGCTCT TCCCCTCTCG ACGTCCCGTG TTCGCAATGC TATGAGACCG CTGACCGGGC TCCCCAGCCC GCCGGTCGCG 14300
14301 CGTGCTTAGC TACGTAGCAG AGGGGCGGTC ACTTGAGGGG CTGCCGCCAC TCGGTGATCT TGCGGCCCAA TGAATCGGCC GCGTTCGAGG CTGCCCGTCC 14400
14401 CACGGCTTTG GTTCACGGTG AAGATCGCAC AGCCGGTGCC GGAAAAGTCC GCGGCACCGA TGTCGGTCAG CAAGACGTTG AAGAGAAACC CCGAGATCAC 14500
14501 CGCCCATGGG ATCGTCATCA ACACCCCAGG CAGCGTCGAC ACCCGCGCCA CGAACCAGCA CTGAAGTAGG TATTCACGCC ACGCGAAAGG CGGCTTGAAC 14600
14601 ATGCACACGG ACGTGTCGAG CGTCATCGCG AAGAAATCGC CCAGCGCGCC CACCGGCCGC AAGACCGGAT CAGCGACCCG ACCGGCCGCC TTGTCGGCCA 14700
14701 CGATTACCAT GGCGCGGCGC ACCAGCTGGA TTCGATGCTG GGCCGTTGGG TGAGGTGGCG CACGCTGGCC CCCCGGACAG GTCGACGATC GGTGACATTG 14800
14801 GTGAGCGTAC GCGGCAGAGA CCGCTGATGT CCATAGCCAA TACGCGATTG CTTGGACAAC TGATCGGTAA ATAGCAATGC AAACTGGCAT ATATTGGCTA 14900
```

Figure 2 continued, page 4 of 4

```
14901 TGATGTATCT TGCTAGTATC CTATAGCGCG GGGCGATGTG CTCTGCTGCC  TTGGCGGCCG ACAGGCGCAT CACCGGTCAA GCCGTTGGCT CGAGTCACGC 15000
15001 TGGCGAGGCA CCACGATCAG GCATCAACAG CGCGCCCGAC GGGCGGTGAT  CGGATGCCGC ATCCTGACCG CCTCGATTCG GGCCCGCCGA CCAGAGCCTT 15100
15101 CGCGACCGGC GAGGTTGCCA CCATGGTCGT CGAAGCAACT TGCTGCTAAC  GAGCCTGTAG TTTTGCCAGC CCCCACTCGC GCTTTGTCTG CAGGTTTTCA 15200
15201 GGCTCAGCGA CGGCTCATGT CGTTGCGCAC GGCGAATTC                                                                    15239
          |  10      |  20      |  30      |  40      |  50      |  60      |  70      |  80      |  90      |  100
```

Figure 3


```
      |  10      |  20      |  30      |  40      |  50      |  60      |  70      |  80      |  90      | 100
    1 GAATTCACTT AGCTAACACC AGTTCTAGCA GCTGTCGGCG CGACTTCTTG TCAGTGCCCG ACGTTATGAT TCGAACATGT TAGCGAATAG CCGGGAGGAG 100
  101 CTTGTCGAGG TCTTTGATGC GCTGGATGCC GAGCTGGACC GCTTGGACGA GGTGTCTTTT GAGGTGTTGA CCACCCCGGA ACGGCTGCGG TCTCTGGAAC 200
  201 GTCTGGAATG CTTGGTGCGC CGGCTACCGG CGGTCGGGCA CACGTTGATC AACCAACTCG ACACCCAAGC CAGCGAGGAA GAACTGGGCG GCACGCTGTG 300
  301 CTGCGCGCTG GCCAACCGGT TACGCATCAC CAAGCCCGAC GCCGCCCTAC GCATCGCCGA CGCCGCCGAT CTCGGACCTC GTCCGAGCAC TCACCGGCGA 400
  401 ACCGCTAGCC CCACAGTTTG ACCGCCACCG CCACCGCCCA ACGCCAGGGC CTGATCGGCG AAGGCGCACA TCAAAGTGAT TCGCGCCCTT TTTCGGCCCA 500
  501 ACCTGCCCGC CGCGGTGGAT GTGTCCAAAC CCGCCAGGCC GCCGAAGCCC GACCTGGCCG CAAACCGCTC AAATATCGTC CCGACGAGCT GGCCCGCTAC 600
  601 GCCCAGCGGG TCATGGACTG GCTACACCCC GACGGCGACC TCACCGACAC CGAACGCGCC CGCAACCGCG GCATCACCCT GAGCAACCAG CAATACGACG 700
  701 GCATGTCACG GCTAAGTGGC TACCTGACCC CCCAAGCGCG GGCCACCTTT GAAGCCGTGC TAGCCAAACT GGCCGCCCCC GGCGCGACCA ACCCCGACGA 800
  801 CCACACCCCG GTCATCGACA CCACCCCCGA TGCGGCCGCC ATCGACCGCG ACACCCGCAG CCAAGCCCAA CGCAACCACG ACGGGCTGCT GGCCGGGCTG 900
  901 CGCGCGCTGA TCGCCTCCGG GGAACTGGGC CAACACAACG GTCTTCCCGT CTCGATCGTG GTCACCACCA CCCTGACCGA CCTGCAAACC GGCGCCGGCA 1000
 1001 AGGGCTTCAC CGGCGGCGGC ACCCTGCTAC CCATGGCCGA TGTGATCCGC ATGACCAGCC ACGCCCACCA CTACTCCCCC GCAAGCGGGA GGTACCCCCA 1100
 1101 GGCGATCTTC GACCACGGCA CACCCCTGGC GCTGTATCAC ACCAAACGCC TAGCCTCCCC GGCCCAGCGG ATCATGCTGT TCGCCAACGA CCGCGGCTGC 1200
 1201 ACCAAACCCG GCTGTGACGC ACCGGCCTAC CACAGCCAAG CCCACCACGT CACCGGCTGG ACCAGCACCG GACGCACCGA CATCACCGAC CTCACCCTGG 1300
 1301 CCTGCGACCC CGACAACCGA CTCGCCGAAA AAGGCTGGAC CACCCGCAAA AACACCCACG GCCACACCGA ATGGCTACCA CCACCCCACC TCGACCACGG 1400
 1401 CCAACCGTGG ACCTGTGAGA TACACTACAC CTGTGCGTGC TGCTGTCTAC CTCCGAATCT CAGAAGACCG CTCCGGCGAA CAGCTCGGCG TGGCCCGCCA 1500
 1501 ACGCGAGGAC TGCCTAAAGC TGTGCGGGCA GCGAAAATGG GTGCCCGTCG AGTACCTCGA CAACGACGTC AGCGCATCAA CCGGCAAGCG CCGCCCCGCC 1600
 1601 TACGAGCAGA TGTTGGCCGA CATCACCGCC GGCAAGATCG CCGCCGTGGT GGCCTGGGAC CTGGACCGGC TCCATCGCCG TCCCATCGAG CTGGAAGCCT 1700
 1701 TCATGTCATT AGCCGACGAG AAGCGGCTGG CCCTGGCCAC CGTCGCCGGC GACGTTGACC TGGCGACACC CCAGGGCCGG CTAGTCGCCC GCCTGAAGGG 1800
 1801 GTCGGTGGCC GCTCACGAAA CCGAGCACAA GAAGGCACGA CAGCGCCGCG CCGCCCGCCA GAAAGCTGAA CGCGGCCACC CCAACTGGTC GAAAGCCTTC 1900
 1901 GGCTACCTGC CCGGCCCCAA CGGTCCCGAA CCCGACCCCC GGACAGCGCC GCTGGTCAAA CAGGCCTACG CCGACATCCT CGCCGGGGCG TCCCTGGGCG 2000
 2001 ACGTGTGCCG CCAGTGGAAC GACGCCGGGG CGTTCACCAT CACCGGCCGC CCGTGGACGA CTACAACGCT GTCGAAATTC TTGCGCAAAC CCCGCAACGC 2100
 2101 CGGACTACGC GCATATAAGG GTGCCCGCTA CGGCCCGGTG GACCGCGACG CGATTGTCGG CAAGGCCCAG TGGTCGCCGC TGGTGGACGA GGCGACGTTC 2200
 2201 TGGGCCGCCC AGGCCGTGCT GGACGCCCCC GGCCGCGCCC CCGGCCGCAA AAGCGTGCGC CGCCACCTGC TGACCGGGCT GGCAGGCTGC GGCAAATGCG 2300
 2301 GCAACCACCT GGCCGGCAGC TACCGCACCG ACGGCCAGGT CGTCTACGTG TGCAAGGCGT GCCACGGGGT GGCCATCCTG GCCGACAACA TCGAACCGAT 2400
 2401 CCTGTATCAC ATCGTGGCCG AGCGGCTGGC CATGCCCGAC GCCGTTGACT TGTTGCGCCG GGAGATTCAC GACGCCGCCG AAGCCGAAAC CATCCGCCTG 2500
 2501 GAACTGGAAA CCCTCTACGG GAGCTGGACA GGCTCGCCGT CGAACGCGCC GAAGGGCTAC TGACCGCGCG CCAGGTGAAG ATCAGCACCG ACATCGTCAA 2600
 2601 CGCCAAGATA ACGAAACTTC AGGCCCGCCA ACAGGATCAG GAACGGCTCC GAGTGTTCGA CGGGATACCG TTGGGAACAC CGCAAGTCGC CGGGATGATA 2700
 2701 GCCGAGCTGT CGCCGGACCG GTTCCGCGCC GTCCTCGACG TCCTCGCTGA AGTCGTTGTC CAGCCGGTCG GCAAGAGCGG CAGGATATTC AATCCCGAAC 2800
 2801 GGGTGCAGGT GAATTGGCGA TGAGCCGGCA CCACAACATC GTGATCGTCT GTGACCACGG CCGCAAAGGC GATGGCCGCA TCGAACACGA GCGCTGCGAT 2900
 2901 CTTGTCGCGC CGATCATTTG GGTCGACGAG ACCCAGGGCT GGTTACCGCA GGCGCCAGCG GTGGCAACAT TACTCGACGA CGACAACCAG CCGCGAGCCG 3000
 3001 TTATTGGCTT GCCGCCCAAC GAGTCTCGCC TACGACCTGA AATGCGCCGC GACGGGTGGG TGCGGCTGCA CTGGGAATTC GCCTGCCTGA GGTACGGCGC 3100
 3101 CGCCGGCGTG CGCACGTGCG AGCAGCGGCC CGTCGGGGTT CGCAACGGCG ACCTGCAAAC ACTGTGCGAG AACGTTCCGC GGCTACTGAC CGGACTGGCC 3200
 3201 GGCAACCCCG ACTACGCACC GGGTTTTGCG GTGCAGTCGG ACGCGGTGGT CGTCGCCATG TGGCTGTGGC GCACGCTCTG CGAAAGCGAC ACGCCGAACA 3300
 3301 AACTACGCGC CACCCCAACG CGTGGTAGCT GCTAGACTCC GACGTAGCCG GCTTCGACTC CGGGGTTTTG GTGTCCCCAA GGAGTCGCAC GTGTCGACCA 3400
 3401 TCTACCATCA TCGCGGCCGC GTAGCCGCAC TGTCTCGTTC CCGCGCATCC GACGATCCCG AGTTCATCGC CGCGAAAACC GATCTCGTTG CCGCGAACAT 3500
 3501 CGCGGACTAC CTCATCCGCA CCCTCGCCGC AGCGCCGCCC CTGACTGACG AGCAGCGCAC CCGGCTGCCC GAGCTGCTGC GCCCCGTGCG GCGGTCAGGC 3600
 3601 GGTGCCCGAT GACGCCGGC GCCGGCGGGT CGCCGCCGAC GCGACGATGC TCGGCCACGG AGGACCGGGC ACCCGCGACA GTCGCCACAC CGTCTAGCGC 3700
 3701 CGATCCTACC GCGTCACGCG CCGTGTCGTG GTGGTCGGTG CACGAGCATG TCGCGCCGGT CCTGGATGCT GCCGGGTCGT GGCCGATGGC CGGCACACCG 3800
 3801 GCCTGGCGTC AGCTCGACGA CGCCGATCCT CGCAAATGGG CCGCGATCTG CGACGCAGCC CGGCACTGGG CTCTGAGGGT AGAGACGTGC CAGGAGGCGA 3900
 3901 TGGCGCAGGC GTCACGTGAC GTATCTGCGG CCGCCGACTG GCCCGGCATC GCCCGCGAGA TCGTCCGACG GCGCGGCGTG TACATCCCGC GGGCGGGGGT 4000
 4001 GGCGTGATGG CCGACATCCC CTACGGCACC GACTATCCCG ACGCCCCCTG GATCGACCGG GACGGGCACG TGCTCATCGA CGACGGTGGC AAACCGACGC 4100
 4101 AAGTTCATCG CGGCCAAGCC CGAATCGCCT ACCGGCTAGC CGAACGTTAC CAGGACAAGC TGCTGCACGT GGCCGGGATC GGCTGGCACT CCTGGGACGG 4200
 4201 CAGACGCTGG GCAGCCGACG ACCGCGGCGA AGCCAAACGT GCAGTGCTGG CAGAGCTGCG CCAAGCGCTC TCAGACAGCC TCAACGACAA GGAATTACGC 4300
 4301 GCCGACGTCC GAAAATGCGA ATCGGCGTCC GGCGTGGCCG GCGTGCTCGA CCTGGCCGCC GCACTGGTAC CATTCGCCGC GACGCTAGCC GACCTCGACA 4400
 4401 GCGACCCGCA CTTGCTCAAC GTCGCGAATG GGACGCTGGA CCTGCACACG CTCAAATTGC GGCCCCACGC GCCCGCTGAC CGCATCACAA AGATATGCCG 4500
```

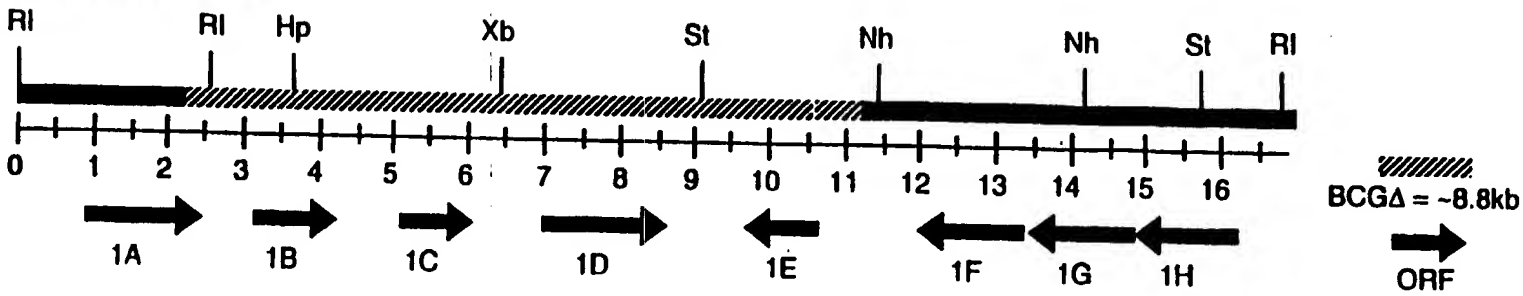
Figure 3 continued, page 2 of 3

```
4501 CGGTGCCTAC CAGTCCGACA CCGAATCGCC TCTCTGGCAA GCGTTCTTGA CCCCGTTCT GCCCGATGAA GGTGTGCGCG GGTTCGTGCA ACGCCTGGCC 4600
4601 GGCGTCGGCC TACTAGGCAC CGTCCGCGAA CATGTCCTGG CGATTCTTAT CGGTGTAGGT GCCAACGGAA AATCTGTGTT CGACAAGGCG ATTCGCTATG 4700
4701 CCCTTGGCGA TTATGCCTGC ACCGCTGAGC CTGACCTTTT CATGCACCGG GAAAACGCTC ACCCAACAGG CGAAATGGAC CTCCGCGGCG TGCGATGGGT 4800
4801 AGCGGTATCC GAGAGCGAAA AAGATCGCCG GCTGGCCGAA TCAACGATAA AACGGCTGAC TGGCGGCGAC GCCATCCGCG CCCGAAAGAT GCGGCAAGAC 4900
4901 TTCGTGGAAT TCGAGTGGTG CCGTTTGAAG TAGTGATTCC TGCCGACGAG CAGGACCGGG AACTGGACGC ACGGTTGCAG TTGGAGGCCG ACAGCATCCT 5000
5001 GTCCTGGGCG GTGGCCGGAT GGAGCGACTA TCAGCGAATC GGACTATCCC AGCCGGACGC GGTGCTCGCG GCAACGTCGA ATTACCGCGA GGACTCCGAC 5100
5101 ACGATAAAGA GGTTCATCGA CGACGAATGC GTCACCAGCT CGCCGGTGCT GAAAGCCACT ACTACGCATC TGTTCGAGGC GTGGCAAAGG TGGCGGGTGC 5200
5201 AAGAAGGCGT ACCCGAAATC TCGCGCAAAG CGTTCGGCCA GTCGCTCGAC ACCCACGGAT ACCCGGTCAC TGACAAGGCC CGTGATGGTC GTTGGCGGGC 5300
5301 CGGAATAGCG GTGAGAGGGG CCGATGATTT CGATGATTAG CACACCTAAC GTGACGCATG TGACGCATTT CCAGGTTCGC CTACGCGCGC GCACGTATGG 5400
5401 CGGTTATACC GCGCAAACGT CACATGCGTC ACGGCCTGCC GTGCCGTTCT GCCCAGGATG CGGTACCTAC CTGGCCGTTC ACGGCCGCCA CCGGGCGGAC 5500
5501 TGTACCGCCA AACCAGCAAA CACCGGCGGT GCCGCATGAC CGCTGTCGCG ATCACCCCGG CATCCGGCGG TCGGCACAGC GTCCGATTCG CCTACGACTC 5600
5601 TGCGATCGTG TCGTTGATCA AGTCCTCGAT CCCCGCCTAT GCCCGCTCCT GGTCCGCGCA CACCCGCTGC TGGTTCATCG ACGCTGACTG GACCCCACTG 5700
5701 CTGGCCGCCG AGCTGCGCTA CCACGGCCAC ACCGTCACCG GACCCGCCGA CCCGGCGCAA CAGCAGTGCA CCGACTGGGC CAAAGCGTTG TTCCGGGCGG 5800
5801 TCCGACCCCA GCGGACACCC GCCGTGTACA GGGCTTTATC CAAAGTGCTG CACCCCGACG CCCCAACCGG ATGCCCGATA CTGCAACAGC AGCTCAATGC 5900
5901 CGCCAGAACC GCACTTACCA ACCCTGCTTG AAAGGACACA AGCCATGGCT GAAACCCCCG ACCACGCCGA ACTGCGGCGA CGAATCGCCG ACATGGCTTT 6000
6001 CAACGCCGAT GTCGGTATGG CGACCTGCAA ACGCTGTGGT GACGCCGTGC CGTACATCAT CCTGCCGAAC CTGCAGACCG GCGAACCCGT CATGGGTGTC 6100
6101 GCCGACAACA AATGGAAGCG CGCGAACTGT CCCGTCGACG TCGGTAAGCC GTGCCCGTTC CTAATCGCCG AGGGTGTCGC CGACAGTACC GACGACACCA 6200
6201 TAGAGGTCGA CCAGTGACCC CGATCAACCG GCCCCTGACC AACGACGAAC GACAACTGAT GCACGAGCTG GCAGTCCAGG TTGTCTGCTC GCAGACGGGT 6300
6301 TGCTCACCCG ATGCGGCGGT CGAAGCACTC GAATCCTTCG CGAAAGACGG AACACTTATC CTCCGCGGCG ACACCGAGAA CGCCTACCTC GAAGCCGGAG 6400
6401 GCAATGTTCT TGTCCATGCC GATCGTGACT GGCTTCCCTT CCACGCGTCG TATCCCGGCA ACGACCCGCT GCGAGACGCC CGACCTATCG AGCAGGACGA 6500
6501 CGACCAGGGG GCGGGTCGC CATCGTGACC AGGCCCAGCC CGGACACCGC CACGGTGCCG GCGCGCATGC ACGCTCATTA CCTAGACTAA AAATTGATGG 6600
6601 GAGGACCGAT GCCAAGACCA CCGAAACCGG CCCGGCTCAA ACTGGTTGAG CGCCGCTCCC CCGGCCGCGA TTCCGGCGGC CGGAAAGTCC CCGAGTCGCC 6700
6701 GAAGTTTATC CGTCAGGCAC CGGATGCCCC GGACTGGCTC GACGCCGAGG CGCTGGCCGA ATGGCGGCGC GTCGCACCGA CTTTGGAGCG GCTTGACCTG 6800
6801 CTCAAACCTG AGGATCGGGC GCTCCTGTCC GCGTACTGCG AGACCTGGTC CGTCTACGTC GCGGCGGTTC AGCGGGTCCG CGCCGAAGGC CTCACAATTA 6900
6901 CCTCACCGAA ATCCGGTGTC GTGCACCGGA ACCCGGCGGT GACGGTTGCG GAGACGGCGC GCATGCATCT GCTGCGCTTG GCCTCCGAGT TTGGCCTGAC 7000
7001 CCCGGCCGCC GAGCAGCGAC TGGCGGTGGC GCCGGGCGAC GACGGCGACG GGCTCAACCC GTTTGCCCCG GACCGGTGAT GACCTTTTGT GTGTGATACA 7100
7101 ATCGAGTTTG GCATCTCGGC ATCCGCTGAC GCCGGCAGT CGCCGCGGGG CGGCTGGAAC CCGGATAGCG GCCGCCATGC GCCACAAGCG ATTCCGCGCG 7200
7201 TTTCTTGCGT CTGCTAGGTG GTGGCCGAAT TTTGAGTAGC ATCCTTTTCC GCATGGCCGA GCTGCGGTCT GGCGAAGGCC GAACCGTGCA CGGCACCATC 7300
7301 GTGCCCTACA ACGAGGCGAC CACCGTCCGC GACTTCGACG GCGAGTTCCA GGAAATGTTC GCTCCTGGCG CTTTTCGGCG CTCCATCGCC GAGCGCGGCC 7400
7401 ACAAATTGAA GCTGCTGGTC TCTCACGACG CTCGAACCCG CTACCCGGTG GGCCGGGCCG TTGAGTTCCG GGAGGAGCCT CACGGCTTGT TCGGGGCGTT 7500
7501 CGAGATTGCG GACACCCCGG ACGGCGACGA GGCTTTGGCG AACGTAAAAG CTGGTGTCGT CGACTCGTTT TCGGTGGGTT TCCGACCGAT CCGGGACCGT 7600
7601 CGCGAAGGGG ATGTGCTGGT GCGCGTCGAA GCGGCGCTGT TAGAGGTTTC CCTAACCGGC GTTCCGGCCT ATTCGGGGGC ACAAATCGCC GGGGTGCGCG 7700
7701 CGGAATCGCT TACAGTCGTT TCCCGTTCGA CAGCCGAAGC CTGGCTGTCC CTACTCGATT GGTGAACAAT CTATGACCGA ATTCGACGAC ATCAAAAACC 7800
7801 TCTCTTTACC TGAAACCCGT GACGCGGCGA AGCAGCTCCT CGACAGTGTC GCCGTGTGAC CTGACCGGTG AGGCGGCGCA GCGTTATTCA GGCGCTGACG 7900
7901 CGCCACGCCG AGGAACTGCG GGCGGAGCAG CGCCGCCGCG GCCGCGAAGC CGAGGAGGAG CTGCGCCGCT ACCGGGCCGG TGAGCTGAGG GTGGTGCCCG 8000
8001 GCGCTCCCAC CGGCGGCGAC GACGGCGACG CGCCGCCGGG CAACTCGTTG CGGGACACCG CGTTTCGCAC ACTGGATTCT TGTGTGCGAG ACGGCCTGAT 8100
8101 GTCGTCGCGG GCGGCGGAGA CCGCCGAAAC CTTGTGCCGC ACCGGGCCGC CGCAGTCCAC CTCGTGGGCG CAGCGCTGGC TGGCGGCCAC CGGCAGCCGC 8200
8201 GACTATTTGG GCGCGTTCGT CAAGCGGGTT TCCAATCCTG TTGCGGGGCA CACGGTTTGG ACCGACCGGG AAGCGGCCGC GTGGCGTGAG GCTGCCGCGG 8300
8301 TGGCCGCCGA GCAGCGAGCG ATGGGCCTGG TGGACACCCA AGGCGGGTTT CTGATCCCGG CGGCGCTGGA CCCGGCGATC CTGCTGTCGG GTGATGGGTC 8400
8401 GACGAACCCG ATTCGGCAGG TGGCGAGGGT GGTGCAAACG ACCTCCGAGA TTTGGCGGGG CGTGACTTCC GAAGGCGCCG AAGCTCGTTG GTACTCCGAA 8500
8501 GCCCAGGAGG TGTCCGACGA TTCGCCAGCG TTGGCCCAGC CGGCGGTGCC GAACTACCGT GGAAGCTGCT GGATTCCGTT CTCCATCGAG CTGGAGGGTG 8600
8601 ACGCGCGCGAG CTTCGTTGGC GAGATCGGCA AGATTCTCGC GGACAGCGTT GAGCAACTGC AGACCGCGGC GTTCGTCAAC GGCTCCGGCA ACGGCGAGCC 8700
8701 CACCGGGTTC GTCAGCGCGC TAACCGGCAC CTCCGATCAG GTGGTCGTCG GCGCGGGGTC AGAAGCGATT GTGGCGGCGG ATGTTTACGC GTTGCAGTCG 8800
8801 GCGCTGCCGC CAAGGTTCCA GGCCAGCGCC GCGTTCGCGG CGAACTTGTC CACCATCAAC ACGTTGCGGC AGCCGGAAAC TTCGAATGGC GCGCTGAAAT 8900
8901 TCCCATCGCT GCACGACAGT CCGCCGATGC TAGCCGGGAA GTCTGTCCTG GAAGTCTCCC ACATGGACAC CGTTGATTCG GCGGTGACAG CGACGAATCA 9000
9001 TCCACTGGTG CTTGGCGACT GGAAGCAATT CCTCATCGTC GACAGAGTTG GGTCCATGGT GGAGTTGGTG CCTCACCTGT TCGGGCCGAA TCGCCGGCCG 9100
9101 ACCGGGCAGC GCGGATTCTT CGCCTGGTTC AGGGTCGGAT CAGATGTGCT GGTGGCAAC GCGTTTCGAG TTCTGAAGGT GGAGACTACC GCGTAGGTAG 9200
9201 GATAGGGCCA GGCGTGGCCG GCCTCTGCTT AGGGGTGCCG GGCCGGCCAC GCCCGCCAAC TCCCCTGCGG GTTGCGTTGT CGATTCGTNN NNNNNNNNN 9300
9301 NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN 9400
9401 NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNCC AAGCCAGAAT 9500
9501 ATCGAGCCTG GCGGCCATGG TCGCCGCCTT CCTGTTGCCG CTGCTTGGCT TTCGGCCGTT CCAGCTCGGC GATCCGGCGG CCAGCGGCGC CATTTGTTTC 9600
9601 TCCGCGAACA GGCGGATTTC TTTGTCGTCG CTGCGGGTTG CGTTGTCGAT TCGTTTGAGC CGCTTGTAGG TGCCGGCGGA GATGCCGAGG GCTGCGCCTA 9700
```

Figure 3 continued, page 3 of 3

```
 9701 CCTCCTTGTC AGTGTGGCGC TGAGACGGCT TTGGTTCCAT GGGACCAAAG  CCGGCATTGG TGATCGATGC ACCGAGGCGA CCACCCTCGC GTTGGCGCTC 9800
 9801 CTTGGCTTTC GGGCGTTCCA GCTCGGCGAT CCGGCGACCA GCGGCGCCAT  TTGTTTCTCC GCGAACCGGC GGATTTCTTT GTCGTCGCTG TGGGTTGCGT 9900
 9901 TGTCGATTCG TTTGAGCCGC CGGTAGGTGC CGGCGGAGAT GCCGAGGGCT  GCGCCGATAG CAGTGTCTGT TTTCGTCGAA TGACGCTCTG ATTCTGGTTT 10000
10001 GTAGCCCATG GGCCCCAAAC CAGAATATCG AGCCTGGCGG CCATGATCCT  GCCCCTCGCG CTGCCGCTGC TTGGCTTTCG GCCGCTCCAG CTCCGCGATC 10100
10101 CGGCGGCCAG CGGCGCCATT TGTTTCTCCG AGATAGCTTC CGGCCCATGG  GCCGGAAGCT ATCCATGCCC CGCCCGTGGG ACCGCCCAGC GTCCTGTTGC 10200
10201 CGCGGTGTTC ACCGTCAGCG CTGGTCTTCC GCTGGGCTTC CGCCGCCAAG  CCCGATCATC CGGCTCCACG GCGGGGTGTC GTCGCCGTCG GGCTCGTCGT 10300
10301 CGCCGGCGAG TCCGAGTTGC CGGTTGATGG CGGCTTGTTC CGGCCCGGAG  CGGCCGGCCA GGATCGCCGG GCCGCACTCG TCGCCGCCGG CGGCGTCGGC 10400
10401 GTGCTCGACG CTGATCCGCA GGAACGCTTC GAGCTCGCCG GTGTGCTCGT  GCCGATTCAA CGCGGCCTGC AGCCGAACGA GCGCTTCCCT CACCTCCAAG 10500
10501 GGCGCGTCCG GGAACCACGC CCGGATCGTC TCGGCCACCT GGTCGCGGTC  GCAGACGGCG CGGTCACCGG TTTCCAGGTC GGTTACCGTC ACCAAGTGGT 10600
10601 TGAAACGTGC TGGTGTGGTG GTCATGGTTG ATCTCCTGGC GTGGAATGTT  CTTCAGCAGT CCACGGCCAA CCCCGCACCA ACACCTTCCA CCACCACGAG 10700
10701 AAGCTGCTAC GCCACAACGA CGAGGACAAC CACGACGATC CGTGAGAATC  GCCGCCCGCG AAGATCTTTG GACATCCCCA CATCGACGTG CGTCCTCGCC 10800
10801 ACCTGGCCAG CACCCGCCCG AACCCGCGAG CTGGCCATTA AGACGAAGTT  GCGATCAAAC CCCTTCGCCA TCAAGCTTTT TGGGCCCGCC TCACCCCAGC 10900
10901 AGGTACTCGG CACGCGTGTT GCCGTCCCAG CGGCGCAAGC CGGCGAACCG  ACCGTCAATC CGCGACGGCC GTCCCGCAAT GCGCAGCGCC CGCCCCAATT 11000
11001 GGTGACCACC GACCCGTCGG GCCAGGGTGA CATGGGCGGT CCACTGACCG  GGCAGGCTGT TGGCCATCGG CGCGGGCGCC AGGTGCGGGC CGCAGAGCCG 11100
11101 GTGCACCTCG GCATGCAGGG CCAAAAGCTC GCTGGTCGGC ACCACCAGCC  GGGTGAACAC GACATTGGCC CGCCCGAACA GCACCGGCGC GCCGATCACG 11200
11201 CAGTCCAGCG GCAGCCGACG GGCAACCGCA CCCAGCGGCT CATCGACCTC  CGGGGCGATC CGTTCGGCCA CCGCCAGCGA CACGTGCGGA CGGCTGGCCG 11300
11301 GCGCCTGGCT GGGTATGCCG GCGGCGGCCA ACCCCGCCCA GATGCGCCGG  ATCGCCGCCT CGGTATCGCT GTCGAAGACC AGCTCGATCG AATGCACCAT 11400
11401 CAGCCGACCA GCCCGGCAAC CCAGTTGCGG TCGAACGCCG CCGCGCTCAT  CGCCGCGAAG TCCCCGGCAT CCAGCGACGC GGCCCCGGCG GGCAGAGCGG 11500
11501 CCCGCACCGT AGCAATGCGC GCCAGCGCGG ACCGATTCGA GGCTGCCACC  AACCCGGGCG GGTCCGGCCA GCTGCCGATC ACCAGCCCTG CACATGAAAC 11600
11601 CTGTTGTGCA GCAAGCGCTT CCAACGTCAA CTTGGTGTGG TTGAGGGTGC  CCAGGTCCGC GGTGACCACC ACCAAAGCCG CGGCGGCCAC GTCGACGGCG 11700
11701 ACATCGCGCA GCGTGACGCC CGGCTCGGCG AGTTCGACCA GCAGCCCGCC  CGCCCCCTCG ACGAGGGTCA ACCGCCCGGG ACGGTCCAGG TCTGCGATCA 11800
11801 GCCGCACGAT CTGATCGCGG GCGGGCAACG CCATCCCGGC GTGTTCGGCG  GCGGCGGCCG GGGCCATCGG CTGCGGATAT CGCGCCAAGC CGGCCAGCTG 11900
11901 GGTCACCCCG GCCAACCGGC CGACCTCGGC GAGGTCGTCG TCACCGCGGG  CGGTGCCGGT CTGAACGGGC TTGCACACCG CCACGTCGAT GCCGGCCTGA 12000
12001 CGTGCGGCCG ACGCCAGCGC CGCGCAGACG ACCGTCTTGC CGACCCCCGT  GCCGGTCCCG GTGACGACCA GGATCGTCAA CGGCGCGCCA CGGCGAGAAC 12100
12101 ATCCGTCAGC ACCCGCCGGG CCAGCTCGAG CTCGCCGGCG TTCAGCGATG  CGCGCGCGGT CAGCCGCAGC CGCGACGTAC CCGCGGGCAC CGTCGGCGGC 12200
12201 CGGAAGCAGC CCACCTTGAC CCCGGCGTCC AGGCAGGCCG CCGCGGCGGC  CACTGCCGAC TCCGGCTCGC CCAGGATCAC CGACACCATC GCCGAGTCCG 12300
12301 GCACCGCAGC CACACCGCAC ATCCGCGCAA GTTCACCAGC GTGGTTGAGC  ACCGCCTGCG ATCGCCACGG CTCCGCCTGC AAGACGCGCA GCGCGGCCCG 12400
12401 TGCGGCACCT TC                                                                                                  12412
        |  10      |  20      |  30      |  40      |  50      |  60      |  70      |  80      |  90      |  100
```

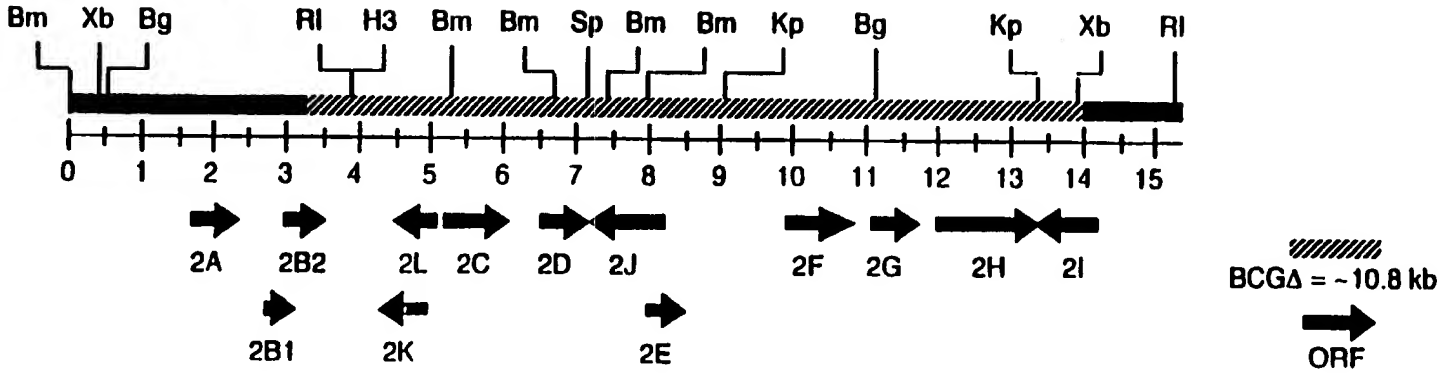
Region 1 (16.9 kb)



ORF	*M. tuberculosis* Codon Usage	ORF Size (base pairs)	Start - Stop (base pairs)	Possible Ribosome Binding Sites	Encoded Protein (max. ~kDa)	Homologies to Predicted Encoded Protein	P value	Homologue Accession #
1A	yes	1542	889 - 2433	AGGA (10)	57			
1B	yes	1071	3130 - 4203 3139 - 4203	GGA (4) GGA (9)	36	*M.leprae* aceA BCG uraA	1.4e-14 3.0e-13	Z46257 U01072
1C	yes	969	5075 - 6046	GAGG (5)	34	*M.tuberculosis* esat6	2.3e-43	X79562
1D	yes	1657	6954 - 8612	none	59			
1E	yes	954	10619 - 9663	GGA (5)	34			
1F	yes	1380	13328 - 11946	AGGA (9)	48			
1G	yes	1386	14823 - 13438 14643 - 13438 14541 - 13438	none GGA (11) AGGAGA (10)	50			
1H	yes	1368	16190 - 14820	GAA (5)	46	*B.subtilis* subtilisin Serine proteases	3.6e-16 4.8e-14	L29506 A08331

Figure 4

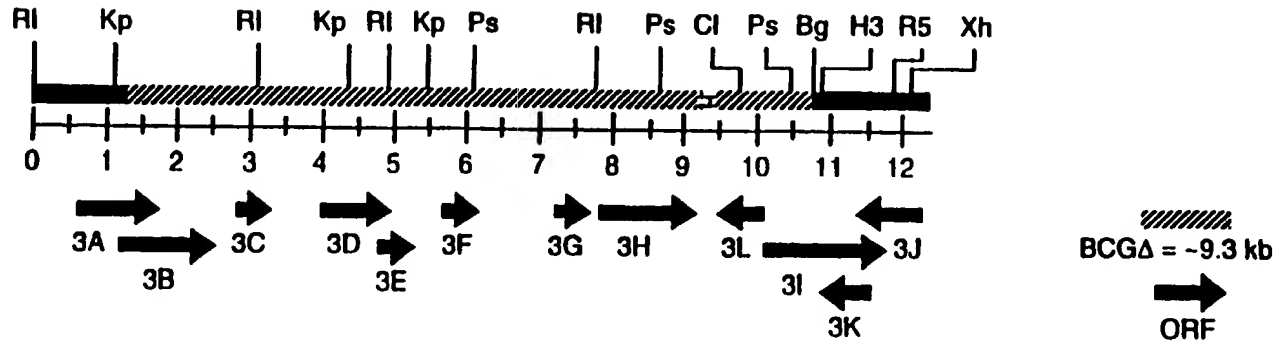
Region 2 (15.3 kb)



ORF	*M. tuberculosis* Codon Usage	ORF Size (base pairs)	Start - Stop (base pairs)	Possible Ribosomal Binding Sites	Encoded Protein (max. -kDa)	Homologies to Predicted Encoded Protein	P value	Homologue Accession #
2A	yes	558	1829 - 2386	AGGGAG (7)	25			
2B1	yes	437	2862 - 3298	AGAA (4)	16			
2B2	yes	588	3003 - 3590	none	34			
2C	yes	948	5187 - 6134 5376 - 6134	AG? (8) GGA (8)	34	*E.coli* iciA lysR family	9.9e-47 <1e-5	P24194
2D	yes	657	6561 - 7217	none	22	*Mleprae* cosmid B1620 ORF Cutinases	1.5e-7 ~4.e-5	U00015 A00975 U03393
2E	yes	522	8036 - 8560	none	19			
2F		966	9941 - 10909	AGGA (11)	37	*S.typhimurium* RNDPR proUVWX	9.9e-146 2.7e-36	X73226 X17445
2G	yes	666	11118 - 11783	AAGA (6)	24	*M.tuberculosis* mpt64	6.7e-141	A30545
2H	yes	1443	11965-13407	AG (10)	51	*E. coli* gabP permease *S. typhimurium* asp permease *T. harzianum* indal gene retroviral receptor	3.1e-11 1.4e-08 4.4e-11 2.5e-09	X65104 U04851 Z22594 X59155
2I	yes	846	14221-13376	GGAAGA (6)	31			
2J	yes	1050	8259 - 7211 7939 - 7211 7931 - 7211	GAG (10) GGAA (8) GGA (9)	35			
2K	yes	666	4992 - 4327	none	25			
2L	yes	597	5117 - 4521	AG (10)	21			

Figure 5

Region 3 (12.4 kb)



ORF	*M. tuberculosis* Codon Usage	ORF Size (base pairs)	Start - Stop (base pairs)	Possible Ribosomal Binding Sites	Encoded Protein (max. ~kDa)	Homologies to Predicted Encoded Protein	P value	Homologue Accession #
3A	yes	1143	613 - 1755	none	42	MTB mce sau3A *M. leprae* cosmid L247	2.9e-64 5.1e-13	X70901 U00021
3B	yes	1347	1214 - 2560		49	Actinophage R4 attP gene *B. subtil.* site spec. recomb. recombinases / invertases	3.0e-05 7.8e-4 8.2e-4	D90361 M29040 K00676 X01805, X07724
3C	yes	513	2820 - 3332	GGA (6)	19			
3D	no no	924	4007 - 4930 4070 - 4930	none GGA (7)	34	*S. coelicolor* phage phi-C31 early region	4.2e-26	X76288
3E	no yes	543	4795 - 5337 4915 - 5337	none GGAA (5)	21	*S. coelicolor* phage phi-C31 pglY pglZ genes	3.2e-11	X76288 L37531
3F	yes	576	5639 - 6214	GA (9)	20			
3G	yes	510	7253 - 7762 7285 - 7762	GAAGG (8)	19			
3H	yes	1330	7868 - 9197	GA (8)	47			
3I	yes	1665	10146 - 11810 10164 - 11810	GGAAG (6)	58	*M. leprae* B1170	6.2e-69	U00010
3J	yes	918	12319 - 11402	GAA (11)	30	*M.leprae* bioDAYB *C. glutamicum* bioD *B.sphaericus* bioDAYB	6.9e-53 2.6e-08 1.4e-05	U00010 D14083 M29292
3K	yes	702	11594 - 10893	AGG.(4)	25	*M. leprae* cosmid B1170	1.0e-81	U00010
3L	yes	660	10147 - ?? 9488	AG (10)	25			

Figure 6

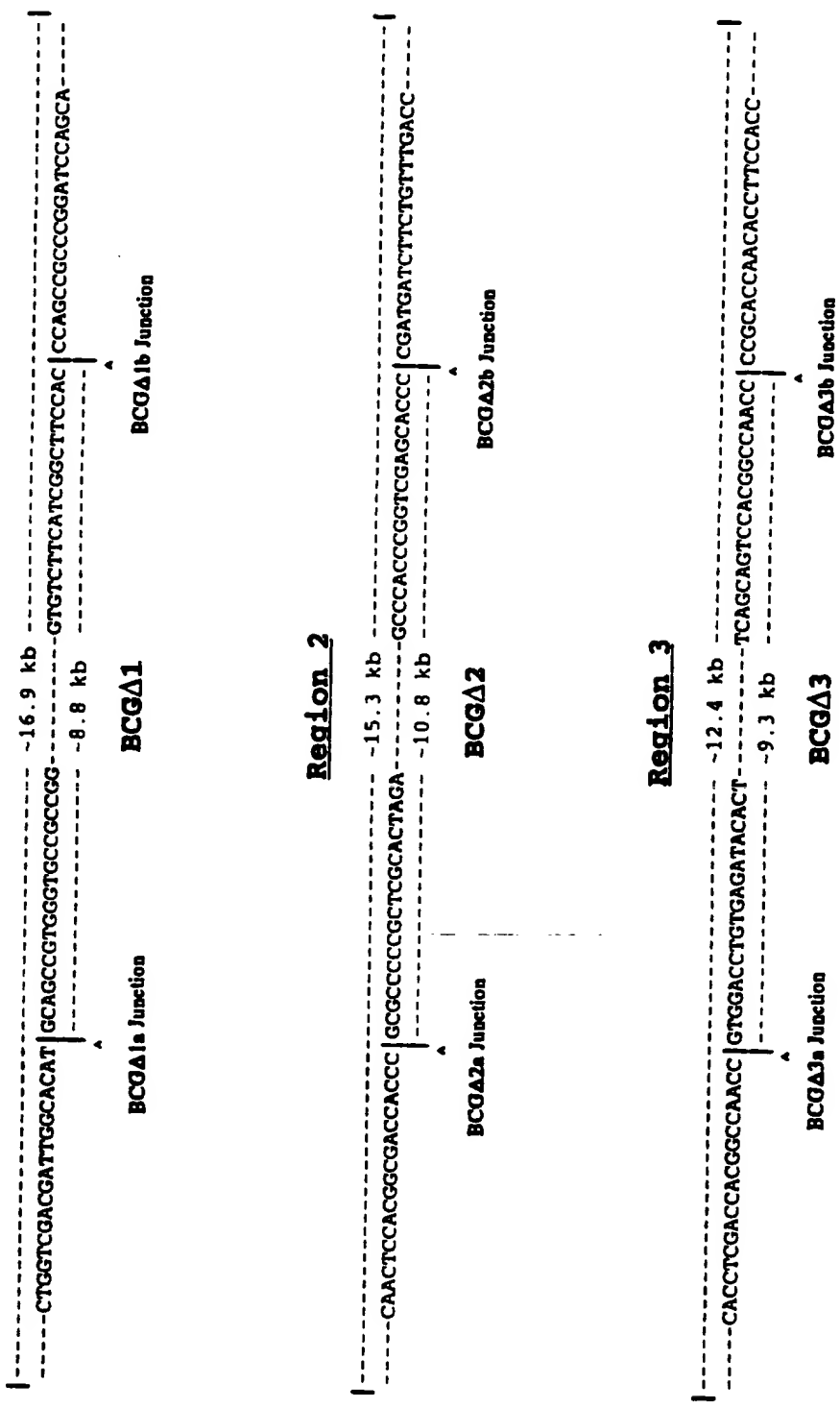

Region 1
-16.9 kb
CTGGTCGACGATTGGCACAT
GCAGCCGTGGGTGCCGCCGG
GTGTCTTCATCGGCTTCCAC
CCAGCCGCCCCGGATCCAGCA
-8.8 kb
BCGΔ1a Junction
BCGΔ1
BCGΔ1b Junction
Region 2
-15.3 kb
CAACTCCACGGCGACCACCC
GCGCCCCCGCTCGCACTAGA
GCCCACCCCGGTCGAGCACCC
CGATGATCTTCTGTTTGACC
-10.8 kb
BCGΔ2a Junction
BCGΔ2
BCGΔ2b Junction
Region 3
-12.4 kb
CACCTCGACCACGGCCAACC
GTGGACCTGTGAGATACACT
TCAGCAGTCCACGGCCAACC
CCGCACCAACACCTTCCACC
-9.3 kb
BCGΔ3a Junction
BCGΔ3
BCGΔ3b Junction

Figure 7

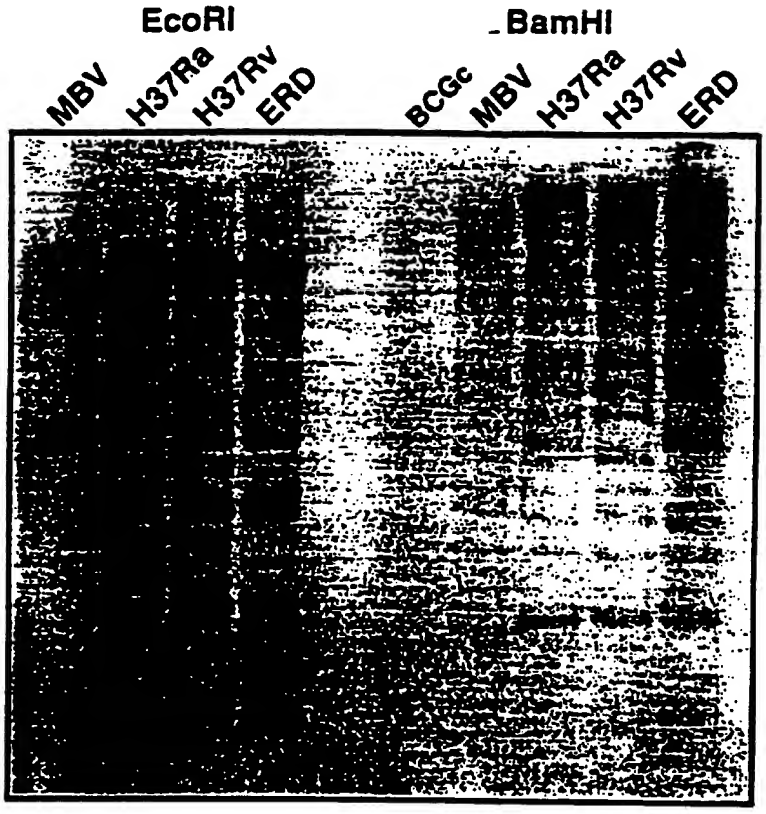


Figure 8

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US96/01938

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :Please See Extra Sheet.
US CL :Please See Extra Sheet.
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6, 7.1, 91.1, 91.2, 240.2, 252.3; 530/300, 350, 387.1, 388.1; 536/22.1, 23.1, 24.3, 24.32, 24.33

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Please See Extra Sheet.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X ---- Y	Infection and Immunity, Volume 61, No. 5, issued May 1993, H. Li et al, "Evidence for absence of the MPB64 gene in some substrains of Mycobacterium bovis BCG", pages 1730-1734, see entire document.	1-10, 16, 17, 24, 25 ----- 18-23
X	JP, 1-247094 (AJINOMOTO ET AL) 02 October 1989, see entire document.	1-7
X	Infection and Immunity, Volume 59, No. 10, issued October 1991, C. Parra et al, "Isolation, characterization and molecular cloning of a specific mycobacterium tuberculosis antigen gene: identification of a species-specific sequence", pages 3411-3417, see entire document.	1-17

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 17 APRIL 1996	Date of mailing of the international search report 29 MAY 1996
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer JEFFREY FREDMAN Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US96/01938

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	Abstracts of the 1994 IDSA Annual Meeting,Clin. Infect. Dis., Volume 19, issued October 1994, R. Frothingham et al, "Sequence based strain differentiation in the Mycobacterium tuberculosis complex, including rapid identification of M. bovis BCG", page 565, see abstract 10.	1-25
X	R. GHERNA et al, "AMERICAN TYPE CULTURE COLLECTION: CATALOGUE OF BACTERIA AND PHAGES", Eighteenth edition, published 1992, pages 202 and 211, see entire document.	11-15
X ---- Y	Infection and Immunity, Volume 62, No. 4, issued April 1994, L. Pascopella et al, "Use of in vivo complementation in Mycobacterium tuberculosis to identify a genomic fragment associated with virulence", pages 1313-1319, see entire document.	1-7, 16-25 ---- 26
Y	Science, Volume 261, issued 10 September 1993, S. Arruda et al, "Cloning of an M. Tuberculosis DNA fragment associated with entry and survival inside cells", pages 1454-1457, see entire document.	1-23
X ---- Y	US,A,5,171,839 (PATARROYO) 15 December 1992, columns 5-10.	1-10 ---- 16-23
Y	Nature, Volume 256, issued 07 August 1975, C. Kohler et al, "Continuous cultures of fused cells secreting antibody of predefined specificity", pages 495-497, see entire document.	10
Y	US,A, 4,683,202 (MULLIS) 28 July 1987, see entire document.	16-22, 24, 25
Y	Genomics, Volume 4, issued 1989, D. Wu et al, "The ligation amplification reaction (LAR) amplification of specific DNA sequences using sequential rounds of template directed ligation", pages 560-569, see figure 2.	16-22, 24, 25
Y	US,A, 4,410,660 (STRAUS) 18 October 1983, columns 14 and 15.	23
Y	Gene, Volume 131, issued 1993, A. Kinger et al, "Identification and cloning of genes differentially expressed in the virulent strain of mycobacterium tuberculosis", pages 113-117, see page 114, column 2.	1-26
X,P	WO,A2,95/17511 (JACOBS ET AL) 29 June 1995, see entire document.	1-26

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US96/01938

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X,E	J. Bacteriol., Volume 178, No. 5, issued March 1996, G. Mahairas et al, "Molecular analysis of genetic differences between mycobacterium bovis BCG and virulent M. bovis", pages 1274-1282, see entire document.	1-26
Y, P	Microbiology, Volume 141, issued 1995, J. Rodriguez et al, "Species-specific identification of mycobacterium bovis by PCR", pages 2131-2138, see entire document.	1-7, 16-22, 24, 25
X ---- Y	Hybridoma, Volume 13, No. 1, issued 1994, A. Arya et al, "Production and characterization of new murine monoclonal antibodies reactive to mycobacterium tuberculosis", pages 21-30, see page 27, table 1.	8-10 ---- 16-23

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US96/01938

A. CLASSIFICATION OF SUBJECT MATTER:
IPC (6):

C12Q 1/68; G01N 33/53; C12P 19/34; C12N 5/10, 1/21; C07K 5/00, 14/00, 16/00; C07H 21/02, 21/04

A. CLASSIFICATION OF SUBJECT MATTER:
US CL :

435/6, 7.1, 91.1, 91.2, 240.2, 252.3; 530/300, 350, 387.1, 388.1; 536/22.1, 23.1, 24.3, 24.32, 24.33

B. FIELDS SEARCHED
Electronic data bases consulted (Name of data base and where practicable terms used):

APS, MEDLINE, BIOSIS, CAPLUS, WPIDS
search terms: mycobacter?, tubercul?, bovis?, BCG, calmette, guerin, DNA, RNA, oligo, nucleic, oligonucleotide, hybridi?, probe, primer, amplif?, PCR, polymerase chain, ligase chain, LCR, attenuat?, immunoassay, antibod?, monoclon?, polyclon?, protein, peptide, antigen, virulenc?, infect?